

Санкт-Петербургский государственный университет

Тотиков Азамат Альбертович

**Реконструкция демографической истории трех популяций обыкновенной
ласки *Mustela nivalis***

Выпускная квалификационная работа
по направлению подготовки «Биология»
основная образовательная программа магистратуры
«Биоинформатика»

Работа выполнена на базе ИМКБ СО РАН и Университета ИТМО

Научный руководитель: с.н.с., к.б.н. Райко Михаил Петрович

Научный консультант: н.с. Кливер Сергей Федорович

Санкт-Петербург, 2021 г.

Оглавление

Введение	3
Обзор литературы	6
Природоохранная биология	6
Природоохранная генетика и геномика	11
Основные методы и задачи природоохранной геномики	13
Способы оценки генетического разнообразия	16
Старые методы низкого разрешения	16
Методы с ограниченной представленностью генома	19
Полногеномный подход	25
Реконструкция демографической истории	31
Методы и их ограничения	32
Значимость таких реконструкций	38
Материалы и методы	43
Оценка качества и фильтрация данных	43
Выравнивание и поиск генетических вариантов	44
Визуализация гетерозиготности	45
Получение и визуализация сетей гаплотипов и филогенетического дерева	45
Реконструкция демографической истории	46
Результаты и обсуждение	47
Оценка качества и фильтрация данных	47
Визуализация покрытия и поиск псевдоаутосомного района (PAR)	52
Поиск генетических вариантов и визуализация гетерозиготности	55
Получение и визуализация сетей гаплотипов и построение филогенетического дерева	58
Реконструкция демографической истории	62
Выводы	66
Благодарности	67
Список литературы	68

Введение

Экологические изменения, вызванные различными антропогенными факторами, представляют серьезную угрозу биологическому разнообразию. Исследования влияния экологических изменений на биологические сообщества проводятся непрерывно, и уже на 2017 год было очевидно, что множество видов позвоночных животных находятся на грани исчезновения [1]. Для предотвращения критической ситуации и смягчения ее последствий, требуются серьезные природоохранные вмешательства, основная цель которых должна заключаться в сохранении и восстановлении биоразнообразия на генетическом, видовом и экосистемном уровнях. Исследования в области генетической изменчивости имеют важное значение для природоохранной биологии. Так, основываясь на анализе генетического разнообразия отдельных популяций, генетика дает представление о различных аспектах природоохранной биологии, включая идентификацию, гибридизацию, родство, историю эволюции, эффективную численность популяции, связь между популяциями, адаптивную генетическую вариативность, локальную адаптацию и инбридинг видов [2]. Генетическая рекомбинация является основой для эволюционных изменений и имеет важное значение для адаптации видов к изменяющемуся климату, среде обитания и различным биотическим факторам. Известно, что низкое генетическое разнообразие популяций позвоночных животных, ставит под угрозу их выживание в долгосрочной перспективе и возможность реинтродукции в случае полного исчезновения или снижения численности в характерных для них местообитаниях [3]. Предрасположенность популяции к вымиранию обычно характеризуется не только низким генетическим потенциалом, но и высоким уровнем инбридинга, что может привести к потере адаптивности, снижению выживания и, следовательно, увеличению риска полного исчезновения [4]. Генетическое разнообразие играет существенную роль в обеспечении устойчивости экосистем, помогает поддерживать функции и стабильность, оказывает такое же воздействие на структуру сообщества и процессы экосистемы, как и видовое разнообразие [5]. Признание все большей важности генетического разнообразия в природоохранной биологии, находит место в качестве одного из основных аспектов сохранения биоразнообразия в рамках Конвенции о биологическом разнообразии (КБР) [6].

Определение эффективного размера численности исследуемой популяции является одним из ключевых в природоохранной геномике, что также отражено в КБР, т.к. общее число особей играет важное значение в определении демографической жизнеспособности популяции. В большинстве исследований по реконструкции демографических историй

различных видов, наблюдаются паттерны, описанные для многих видов млекопитающих по всей Евразии. Эти паттерны часто интерпретируются как последствия глобальных климатических изменений, длящихся до сих пор. Климатические циклы могут приводить к постепенному сокращению численности особей в отдельных популяциях, последующей реколонизации новых местообитаний и адаптации под новые условия среды. В результате на графиках демографической истории мы наблюдаем значительные колебания у многих популяций различных видов млекопитающих. Одним из таких примеров являются белые медведи *Ursus maritimus*, известные своей адаптацией к экстремальным условиям Арктики, которые стали символом угрозы глобальных климатических изменений для биоразнообразия. В период оледенения, численность бурых медведей *Ursus arctos* стала сокращаться, а белых - наоборот, возрасть до определенного предела, а затем произошла смена тренда [7]. Иберийская рысь *Lynx pardinus* также является хорошим примером, когда геномные исследования исчезающих видов восстанавливают динамику численности за прошедшие тысячелетия, выявляют закономерности геномной эрозии, которые могут ограничивать их жизнеспособность, и предлагают инструменты для их эффективного сохранения. Иберийская рысь - это самая исчезающая кошка из семейства кошачьих и уникальный пример вида, находящегося на грани исчезновения. Геномные особенности иберийской рыси, могут препятствовать краткосрочной и долгосрочной жизнеспособности из-за снижения адаптивного потенциала [8].

Одним из наиболее распространенных видов хищников в Северном полушарии является обыкновенная ласка *Mustela nivalis*. Ее ареал охватывает большую часть Палеарктики: Европу, Северную Африку, Северную Азию, включая Японские острова, и Северную Америку (рис. 1). На протяжении всего своего ареала, *M. nivalis* имеет значительные вариации в размерах и пропорциях тела и черепа [9]. Вероятно, что несколько признаков *M. nivalis*, включая экологический статус, всестороннее распространение и большие географические вариации, возникли в результате адаптации к различным условиям окружающей среды. Важно отметить, что в целом, тенденция численности *M. nivalis* хоть и считается устойчивой (согласно Международному союзу охраны природы), отмечаются сокращения численности в некоторых популяциях, что может быть связано, например, с климатическими изменениями, инбредной депрессией, потерей адаптивных особенностей к быстро изменяющимся средам их обитаний под влиянием различных факторов [10]. Демографическая история популяций *M. nivalis*, произошедшая в результате постледниковой реколонизации из разных, а в некоторых случаях и из нескольких участков земной поверхности - рефугиумов, где вид пережил неблагоприятный для него период

геологического времени, в течение которого на больших пространствах этот вид исчезал, представляет большой интерес [10]. Точное число подвидов до сих пор неизвестно. Генетическое разнообразие ранее оценивалось только на уровне нескольких митохондриальных и ядерных маркеров [11]. Таким образом, цель данной работы заключается в сравнении на полногеномном уровне и реконструкции демографической истории трех особей *M. nivalis* из отдельных популяций. Для достижения поставленной цели были определены следующие основные задачи:

1. Оценка качества и фильтрация используемых данных.
2. Выявление и фильтрация генетических вариантов.
3. Визуализация гетерозиготности.
4. Реконструкция демографической истории *Mustela nivalis*.
5. Сравнение полученных результатов реконструкции с демографическими историями других видов.



Рисунок 1. Ареал *Mustela nivalis*, с указанием глобального статуса вида. На карте отмечены точки исследуемых условных популяций: 1 - Польская популяция (Польша, г. Краков), 2 - Новосибирская популяция (Россия, г. Новосибирск), 3 - Дальневосточная популяция (Россия, Сихотэ-Алинский заповедник).

Обзор литературы

Природоохранная биология

Природоохранная биология вобрала в себя весь существующий комплекс мер, направленных на поддержание и восстановление различных биологических видов, а также изучение эволюционных процессов, прямо влияющих на генетическое, популяционное, видовое и экосистемное разнообразие. Термин и концепция «Природоохранная биология», как совершенно новая область научных исследований, впервые возникли в 1978 году в результате созыва «Первой международной конференции по исследованиям в области природоохранной биологии», проведенной в Калифорнийском университете в Сан-Диего. Повестку созыва составляла озабоченность научного сообщества по поводу вырубки тропических лесов, сокращения и вымирания биологических видов и подрыва генетического разнообразия внутри различных популяций [12]. Результатом проведенной конференции стала разработка мер, направленных на интеграцию теоретических основ экологии и эволюционной генетики с одной стороны, и соответствующих политических аспектов и существующей практики сохранения видов, с другой [13]. Позднее, междисциплинарность, составляющая основу для природоохранной биологии, привела к формированию новых направлений, включая природоохранные социологию, поведение и физиологию, что стимулировало развитие природоохранной генетики [14].

Применяемые природоохранные меры для сохранения, рассматриваются в контексте двух важных аспектов: систематического природоохранного планирования и физиологического сохранения. Первый включает эффективный способ поиска и определения наиболее действенных типов проектирования заповедников для поддержания приоритетных ценностей биоразнообразия и работы с сообществами в поддержку локальных экосистем [15]. Второй является интегративной научной дисциплиной, применяющая различные физиологические концепции, инструменты и знания для изучения биологического разнообразия и его экологических последствий, а также понимание и прогнозирование того, как организмы, популяции и экосистемы реагируют на изменения окружающей среды и стрессовые факторы. Физиология рассматривается в максимально широком смысле для включения функциональных реакций на всех уровнях живой материи, а сохранение включает разработку и уточнение стратегий восстановления популяций и экосистем [14]. Подходы к сохранению также могут быть классифицированы как сохранение *in-situ*, которое обеспечивает защиту исчезающего вида в его естественной среде обитания, или как

сохранение *ex-situ*, которое происходит вне естественной среды обитания, например, в резервациях или геномных банках, в условиях, когда жизнеспособные популяции могут отсутствовать в естественной среде обитания [16].

Природоохранная биология имеет определенные проблемы в измерении темпов сокращения численности биологических видов. Скорость фонового вымирания в начале 1990-х годов оценивали как исчезновение одного вида каждые несколько лет [17]. На данный момент, фактические темпы вымирания оцениваются на порядок выше [18]. Измерение сокращения численности осложняется еще и тем, что многие биологические виды не были описаны, а природоохранный статус еще большего количества не был оценен. Значительный вклад в повсеместное понимание ситуации с биоразнообразием, внес созданный в 1964 году Международный союз охраны природы (МСОП) «Красная книга видов, находящихся под угрозой исчезновения», который вобрал в себя наиболее полную информацию о глобальном статусе риска исчезновения видов животных, грибов и растений. Согласно имеющимся данным МСОП, на сегодняшний день оценены многие группы видов, включая млекопитающих, земноводных, птиц, рифообразующих кораллов и хвойных пород. В настоящее время в Красную книгу МСОП включено более 128.500 видов, из которых более 35 500 видов находятся под угрозой исчезновения, в том числе 40% амфибий, 34% хвойных деревьев, 33% рифовых кораллов, 26% млекопитающих и 14% птиц [19]. Общее число видов, включенных в Красную книгу МСОП, ежегодно растет (рис. 2).

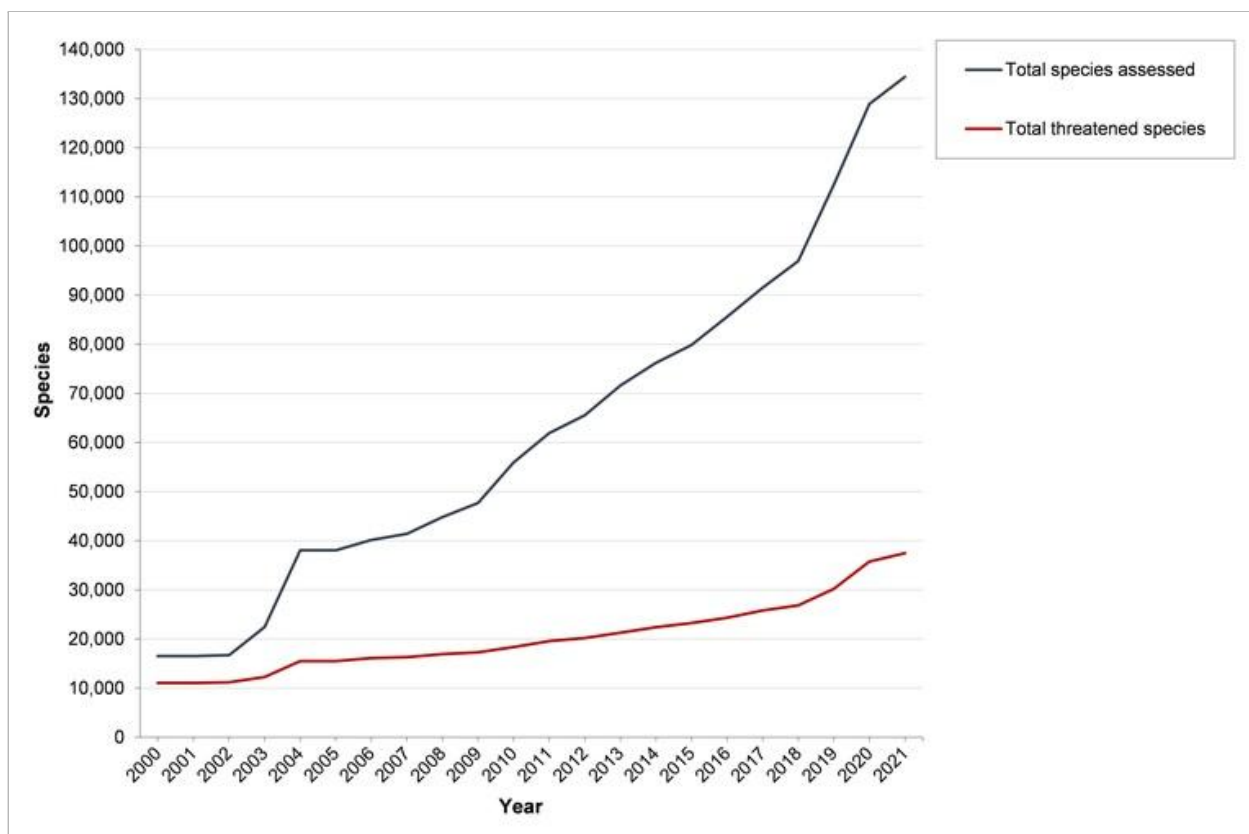


Рисунок 2. Динамика числа видов, включенных в Красную книгу МСОП [19]. Синяя кривая соответствует числу оцененных видов (данные с 2000 года по 2021 год), красная - количество угрожаемых видов, соответственно.

Красная книга МСОП служит международным природоохранным инструментом для выявления видов, наиболее нуждающихся в природоохранном вмешательстве, и для составления глобального индекса состояния биоразнообразия. Однако текущее, шестое по счету массовое вымирание видов, является острым кризисом биоразнообразия, требующего гораздо более активных действий, чем приоритетное внимание к редким, эндемичным или находящимся под угрозой исчезновения видам [1]. Озабоченность утратой биоразнообразия охватывает более широкий спектр проблем, в котором рассматриваются экологические процессы, такие как миграция, и целостное изучение биоразнообразия на уровнях, выходящих за пределы видов, включая генетическое, популяционное и экосистемное разнообразие [20]. Обширные, систематические и высокие темпы снижения биоразнообразия угрожают устойчивости экосистем, регулирующей сложными эволюционными процессами генетического разнообразия. Хотя Красная книга МСОП дает хорошее представление о текущем состоянии видов, ее не следует интерпретировать как полную и достаточную оценку мирового биоразнообразия. При рассмотрении количества видов в каждой категории

Красной книги и пропорции видов находящихся под угрозой исчезновения, важно учитывать, что не все таксономические группы были оценены полностью. Разработанный Индекс Красной книги (ИКК), основанный только на реальных данных снижения и повышения численности различных групп видов, показывает тенденции к ускорению темпов утраты биоразнообразия (рис. 3). Сохранность видов широко используется в управлении природоохранной деятельностью, но некоторые исследователи подчеркивают, что именно обычные виды являются основным источником эксплуатации и изменения местообитаний человечеством. Более того, обычные виды часто недооцениваются, несмотря на их роль в качестве основного источника экосистемной стабильности [21,22]. Важно отметить, что еще и виды-космополиты, имеющие общий глобальный статус как «не угрожаемые виды», зачастую не входят в зону интересов природоохранной биологии, хоть и являются одними из наиболее уязвимых групп животных. Для некоторых видов-космополитов, например, обыкновенной ласки *Mustela nivalis*, отмечается сокращение численности в некоторых популяциях, а точное количество подвидов до сих пор остается неизвестным [10]. Необходимы локальные оценки на уровне подвидов и отдельных популяций, кроме того, для таких широкораспространенных видов также необходимо провести уточнение таксономической и популяционной структуры. Видовое и генетическое разнообразие вносит свой вклад в стабильность экосистем, экосистемных процессов, эволюционную адаптивность и биомассу, а сохранение биоразнообразия критически необходимо для снижения темпов вымирания видов.

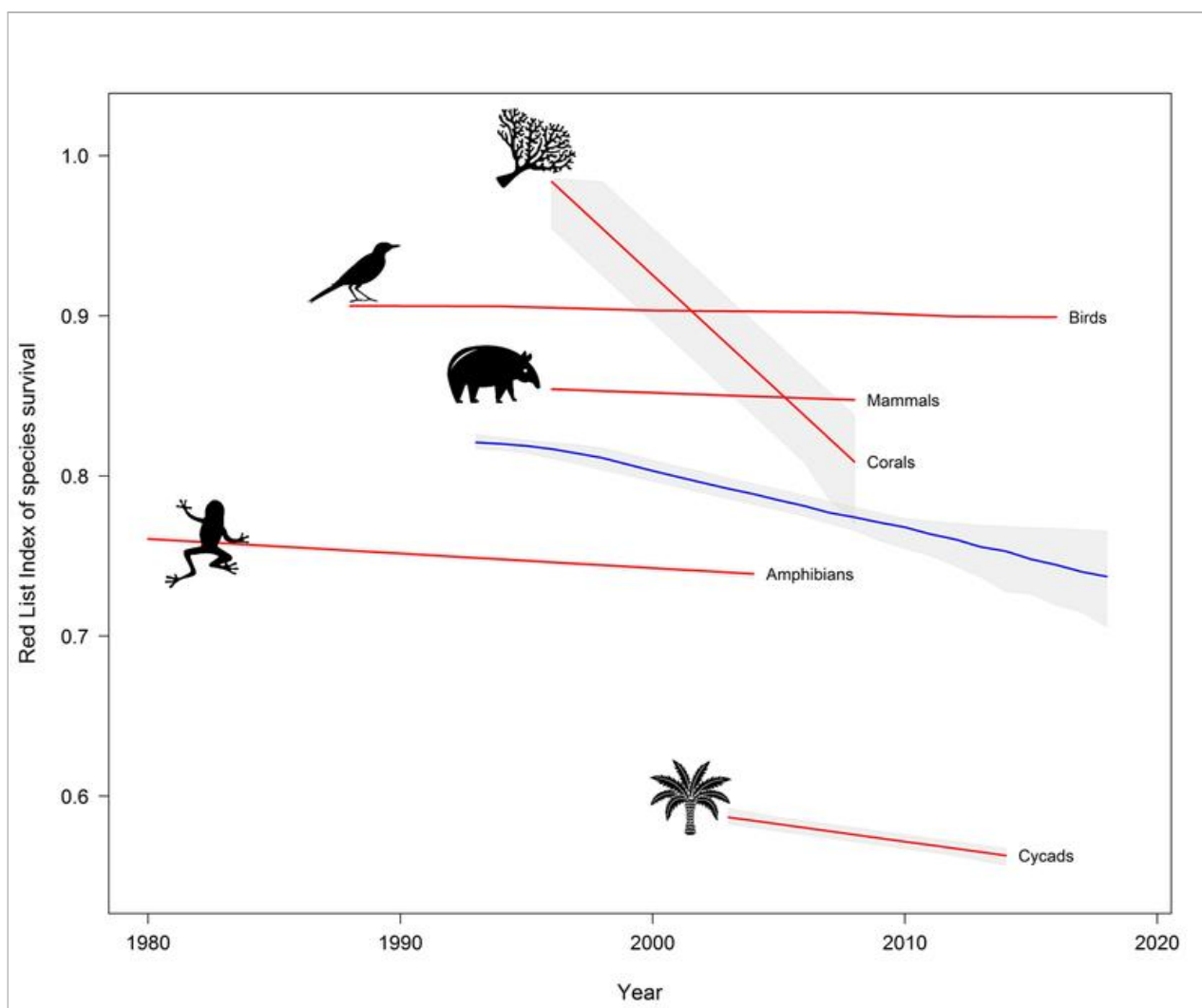


Рисунок 3. Индекс Красной книги (ИКК) выживания видов млекопитающих, птиц, земноводных, рифообразующих кораллов и саговниковых [19]. Синяя линия показывает общий ИКК для всех таксонов. Значение ИКК, равное 1.0, соответствует всем видам, относящимся к категории не угрожаемых (т.е. не ожидается, что они станут вымершими в ближайшем будущем). Значение ИКК, равное 0, означает, что все виды вымерли. Постоянное значение ИКК с течением времени указывает на то, что общий риск исчезновения для группы не изменился. Если бы темпы утраты биоразнообразия сокращались, ИКК демонстрировал бы тенденцию к росту.

Природоохранные генетика и геномика

Поскольку все большее внимание уделяется глобальному изменению климата и высокой скорости снижения биоразнообразия, крайне важно понять изменения и проблемы, с которыми сталкиваются популяции диких животных, и использовать наиболее эффективные инструменты, доступные в настоящее время для сохранения биологических видов [23]. Основные вопросы сохранения дикой природы включают определение структуры и оценку численности популяций, а также взаимодействие между ними, выявление гибридизации, оценку генетического разнообразия, оценку потенциала популяции для сохранения и адаптации к изменениям окружающей среды, а также необходимость понимания факторов, влияющих на этот потенциал. Генетические данные могут послужить источником информации по всем этим вопросам и предоставить возможность для разработки стратегий управления, направленных на их решение. Благодаря использованию методов генетики, как инструментов в изучении природоохранной биологии, возникла природоохранная генетика, как новая научная дисциплина. Природоохранная генетика дала важную информацию о динамике исчезающих популяций. В частности, это привело к эмпирическому пониманию того, как процесс инбридинга и увеличения генетического дрейфа приводит к потере генетической изменчивости в небольших изолированных популяциях [24]. Это способствовало широкому применению молекулярных методов в природоохранной биологии, таких как анализ генетической изменчивости в аллозимных локусах и на уровне ДНК: отдельных локусов в органелльной и ядерной ДНК, микросателлитов и некоторых других. Кроме того, природоохранная генетика внесла важный вклад в интерпретацию структуры метапопуляций, потока генов, демографии, эффективной численности и эволюционной истории [25,26].

Геномные подходы стали важной частью природоохранной биологии и используются для изучения определенных процессов в природоохранной биологии, например, динамики изменения численности популяций, разрешения тесно связанных вопросов о адаптивных способностях, сохранения генетической изменчивости и идентификации различных экотопов, а также непосредственных действий по сохранению и управлению популяциями и их средой обитания. Применение данных природоохранной геномики может вызвать определенные затруднения, по сравнению с данными природоохранной генетики. Ключевая разница между ними заключается в объеме - геномные данные представляют собой информацию по всему геному для одной особи, поэтому возникает необходимость в их тщательной обработке с последующим эффективным применением с конкретной целью для

природоохранных действий, например, такие данные могут иметь решающее значение для оценки жизнеспособности популяции. Не так давно технологический прогресс и связанное с ним падение стоимости высокопроизводительного секвенирования, дали возможность раскрыть потенциал полногеномных методов в области природоохранной геномики, позволив использовать полногеномные данные практически для любой популяции, включая популяции редких, трудно изучаемых и немодельных видов [2,27].

Природоохранная генетика и ее дальнейшее развитие в геномику сильно интегрированы и взаимозависимы. Подтверждением этому могут служить исследования однонуклеотидных полиморфизмов (SNP) и структуры популяции белой куропатки *Lagopus lagopus* и европейского мерлуза *Merluccius merluccius* [28,29]. Оба исследования начались с секвенирования генов и обнаружения SNP для выяснения структуры популяции. В исследовании *L. lagopus* использовалась геномная информация и было секвенировано несколько генов-кандидатов. В исследовании *M. merluccius* использовался RNA-seq (метод определения первичной структуры молекул РНК, представляющий собой высокочувствительный и точный инструмент для изучения транскриптома), для обнаружения SNP с помощью более современных методов NGS. В первом исследовании было обнаружено 127 SNP, а во втором - 381. Эти исследования очень похожи, поскольку оба посвящены важным вопросам природоохранной биологии, демонстрируя, что методы природоохранной генетики и геномики не исключают друг друга. Природоохранная геномика имеет значение для определения локусов генома, отвечающих за локальную адаптацию, а также предоставляет инструменты для дополнительных исследований, таких как оценка эффективного размера численности популяции [30]. Геномные данные активно используются в контексте природоохранной биологии и при оценке демографической истории. Здесь старые генетические методы терпят неудачу, потому что при небольшом количестве маркеров, трудно отделить демографические события, например, узкие места в размере популяции, от отбора. Однако геномные данные, даже от одной особи, могут предоставить информацию о динамике популяции исчезающих видов в прошлом [31]. Данная информация позволяет определять, находились ли вымирающие виды под угрозой исчезновения и были ли узкие места также в их прошлой эволюционной истории, или же их нынешний статус угрозы является следствием того, что происходит в настоящее время. В свою очередь, природоохранная генетика предоставила дополнение, и, в некотором смысле, рентабельные альтернативы дорогостоящему и трудоемкому отслеживанию особей при миграции и расселении [32].

Основные методы и задачи природоохранной геномики

Традиционная генетика сохранения дикой природы опирается на такие методы, как генотипирование микросателлитов и секвенирование отдельных ядерных локусов или митохондриальной ДНК, для получения обширных знаний о различных популяциях [25]. Однако эти методы позволяют получить данные лишь по ограниченному числу генетических маркеров. Совершенствование технологий секвенирования следующего поколения (NGS) привело к увеличению числа методов исследования геномики популяций, все из которых потенциально способны обеспечить получение точных генетических данных по геному не только одной особи, но и всей популяции [33]. Ряд методов, используемых популяционной геномикой, позволяют получить данные о последовательностях в относительно небольшой части генома, например, транскриптом или предварительно отобранный набор локусов [34]. Методы секвенирования с использованием рестрикции (RADseq - Restriction site associated DNA sequencing) обеспечивают получение данных о последовательностях множества отдельных локусов, распределенных по всему геному. При этом расположение локусов определяется выбором рестриктаз и встречаемостью соответствующих сайтов рестрикции в конкретном геноме [35]. Наконец, полногеномное секвенирование (WGS) позволяет получить данные практически для всего генома, за исключением сложных для секвенирования и сборки локусов (конститутивного гетерохроматина, GC-богатых участков и т.д.), и это становится все более актуальным для большинства таксонов [2].

Внутренние границы между генетикой и геномикой, а также вопрос о том, имеет ли смысл их проводить, являются предметом для обсуждения. Огромное увеличение объема данных, предоставляемых методами геномики, может позволить решить новые вопросы, такие как обнаружение генов, связанных с важной функциональной способностью и высокой степенью адаптивности, которые ранее не решались традиционными методами [36]. Локализация генетических маркеров в геномных сборках хромосомного уровня, может дать важную информацию о физическом сцеплении и рекомбинации, и связывает генетические маркеры напрямую с генами-кандидатами. В подходе «геномики в широком смысле» инструменты высокопроизводительного секвенирования могут использоваться для решения вопросов, которые уже можно было решить с помощью традиционных генетических методов [37]. Преимущество использования новых методов заключается в увеличении статистической мощности и разрешающей способности за счет большего количества маркеров, а во многих случаях - в повышении эффективности и рентабельности [38].

Прогресс применения методов геномики для понимания популяций различных животных выявил несколько общих тенденций. Во-первых, все методы, от традиционных инструментов генетики, до широкого спектра подходов секвенирования, продолжают играть важную роль в природоохранной биологии. Выбор наиболее эффективного подхода в каждом конкретном научном исследовании зависит от многих факторов, включая доступные ресурсы и необходимые данные [38]. Во-вторых, исследования природоохранной геномики все чаще позволяют решать множество научных вопросов на основе единого набора геномных данных. Например, геномные данные могут позволить оценить структуру популяции с точки зрения как нейтральной, так и адаптивной связи, с различными последствиями для дальнейших природоохранных мер [39]. Данные WGS от относительно небольшого числа видов, могут предоставить информацию в различных временных масштабах, от демографической истории и филогеографических отношений между широко разделенными популяциями за последние два миллиона лет, до инбридинга в течение последнего столетия [40]. Частично, это является результатом новых аналитических подходов, ставших возможными благодаря наборам геномных данных, таких как демографическая реконструкция и ряды гомозиготности. В-третьих, некоторые подходы природоохранной биологии сочетают в себе несколько методов популяционной генетики и геномики. Например, для многих задач природоохранной генетики требуется последовательное генотипирование набора генетических маркеров, что необходимо при долгосрочном мониторинге конкретной популяции. Методами NGS можно эффективно получить большой объем биологических данных, при обработке которых можно подобрать высокооптимизированный набор (панель) маркерных локусов для определенных целей, таких как анализ происхождения, определение принадлежности к конкретной популяции или мониторинг адаптивных локусов [27,41,42]. Такие панели, как правило, включают небольшое количество локусов, от единиц, до нескольких десятков (на порядки меньше, чем набор геномных данных, на которых они основаны), следовательно, большая часть генома исключается из анализа, поэтому их нельзя считать полногеномными в современном понимании этого слова. Тем не менее, когда выбранная панель маркерных локусов разрабатывается на основе полногеномного набора данных для включения адаптивно значимых локусов, она может ответить на вопросы об адаптивных вариациях в популяциях диких животных, которые ранее были трудноразрешимыми с помощью традиционных генетических исследований. Полногеномные данные позволяют получать репрезентативные панели маркерных локусов, например, в случае сложной зоны гибридизации, не только при скрещивании F1, но и F2 и т.д., микросателлиты должны генотипировать каждую хромосому в лучшем случае. На примере семейства *Mustelidae*, межвидовая амплификация

полиморфных микросателлитных локусов демонстрирует большую вариабельность. В исследовании с использованием 13 праймеров микросателлитных локусов американской куницы *Martes americana* для амплификации последовательностей у других видов семейства *Mustelidae* - ильки (*Pekania pennanti*), американской норки (*Neovison vison*), канадской выдры (*Lutra canadensis*), росوماхи (*Gulo gulo*) и американского барсука (*Taxidea taxus*), показали успешность амплификации в 76.9%, 46.2%, 30.8%, 76.9% и 38.5% соответственно [43]. Исследование изменчивости 33 полиморфных микросателлитных локусов каменной куницы (*Martes foina*), *Martes americana* и *Neovison vison* в трех референсных популяциях соболя (*Martes zibellina*), помогло определить 12 полиморфных локусов для индивидуализации *Martes zibellina*, путем секвенировали всех повторяющихся мотивов каждого локуса, для достоверности правильного генотипирования [44].

Для большого числа животных доступны референсные сборки генома, которые обеспечивают ряд преимуществ, включая точную идентификацию локусов, определение связи генетических маркеров с генами-кандидатами, а также анализ на основе гаплотипов или другие анализы, невозможные в противном случае [45]. В ситуации, когда референсный геном недоступен для определенного вида, используют имеющийся референсный геном близкородственного вида для дальнейшего выравнивания данных коротко считываемых последовательностей, а также этот геном может служить основой для создания референсной сборки генома для основных видов путем скаффолдинга по референсу. Так, например, полногеномные данные снежного барса были выровнены на сборку генома тигра, т.к. оба относятся к роду *Panthera* [46]. В данной работе черновая полногеномная сборка обыкновенной ласки (*Mustela nivalis*) была собрана до псевдохромосом на основе выравнивания на референсную сборку генома горностая (*Mustela erminea*). Оба данных вида также относятся к одному и тому же роду.

Способы оценки генетического разнообразия

Старые методы низкого разрешения

Одним из основных аспектов исследования в природоохранной биологии, является определение размера изучаемой популяции, который можно решить с помощью инструментов популяционной геномики. Общее число особей может быть ключевым фактором в определении демографической жизнеспособности популяций и в определении различных природоохранных действий, таких как нормы вылова, основанные на количестве взрослых особей, темпах увеличения и динамике популяции в целом. Маркерные панели, являющиеся инструментами природоохранной генетики и разработанные для индивидуальной идентификации, могут использоваться в исследованиях природоохранной геномики для оценки плотности и размера популяции, включая неинвазивные образцы, такие как помет и шерсть [27,47]. Инструменты геномики могут предоставить эффективные методы для создания таких панелей маркеров на основе строгой фильтрации гораздо большего набора локусов.

Определение размера популяции имеет решающее значение не только для реконструкции демографической истории, но и для определения генетического разнообразия. Это находит отражение в эффективном размере популяции (N_e), определяемом как размер идеальной, панмиктической популяции, которая в результате генетического дрейфа подвергнется той же потере генетической изменчивости, что и наблюдаемая популяция. N_e обычно меньше, чем наблюдаемая численность популяции (N_c), в силу ряда факторов, общих для естественных популяций, включая колебания численности популяции, дисперсию в репродуктивном успехе и перекрытие поколений, хотя соотношение N_e/N_c сильно варьируется [48]. N_e влияет на вероятность накопления патогенных генетических вариантов, депрессию генов в период инбридинга и способность популяции адаптироваться к изменениям окружающей среды или заболеваниям - важным факторам в популяциях дикой природы, численность которых снижается или которые проходят так называемое «bottleneck» - «узкое место» или «бутылочные горлышко».

Микросателлитная ДНК, также известная как простые повторы последовательностей (SSR) или короткие tandemные повторы (STR), являются распространенными повторяющимися последовательностями в геномах эукариот. Обычно они состоят из tandemно повторяющихся мотивов от 1 до 6 п.н. [49]. Фланкирующие области повторяющихся последовательностей в микросателлитных локусах в основном

консервативны, а мотивы сильно варьируются у разных видов и даже у разных особей одного и того же вида. Таким образом, возможно конструировать специфические праймеры на основе консервативных последовательностей и амплифицировать последовательности повторов с помощью полимеразной цепной реакции (ПЦР), затем генетические полиморфизмы могут быть обнаружены с помощью электрофореза [49]. Преимущества и недостатки маркеров STR были определены многими исследователями [50–52]. К преимуществам обычно относят необходимость небольшого количества матричной ДНК (от 10 до 100 нг.), невысокие требования к качеству и целостности образцов, высокий полиморфизм, кодоминантность маркеров, относительно высокую точность и воспроизводимость, а также отмечается, что различные микросателлиты могут быть мультиплексированы при проведении ПЦР. К недостаткам относят трудоемкость и дороговизну разработки классическими методами, вероятность ошибочной классификации гетерозигот как гомозигот, возникновение нулевых аллелей и последующее затруднение точной оценки полиморфизмов. Основная модель мутаций в значительной степени неизвестна и микросателлитные маркеры действительно помогают определить нейтральное биоразнообразие, но при этом не предоставляют информацию о функциональных характеристиках биоразнообразия. Несмотря на эти недостатки, микросателлитные маркеры являются популярными ядерными ДНК-маркерами для исследования межвидовой и внутривидовой генетической изменчивости.

Разработка и последующее применение микросателлитов в природоохранной генетике стало крупным достижением, т.к. это позволило избежать ограничения в описании генетической изменчивости исследуемого вида [53]. Скрининг многих микросателлитных локусов и возможность выбора набора сильно варьируемых локусов, позволяет исследовать большое количество полиморфизмов практически у любых видов. Например, наличие множества аллелей в микросателлитных локусах предоставило ценную возможность проверить недавнее сокращение размеров популяции, вызванное антропогенными факторами [54]. Хотя оценка генетической изменчивости в аллозимных локусах позволила выявить многие виды с низким уровнем генетической изменчивости (например, северного морского слона (*Mirounga angustirostris*) и гепарда (*Acinonyx jubatus*)), но, тем не менее, было трудно определить, было ли такое небольшое количество генетических вариаций вызвано недавним антропогенным воздействием или такие виды имели исторически небольшой размер популяции. Определение низкого уровня генетического разнообразия имеет решающее значение, поскольку популяции, переживающие недавние узкие места, с большей вероятностью будут содержать патогенные рецессивные аллели с высокой частотой мутации

[55]. Бутылочные горлышки приводят к потере гетерозиготности и аллельного разнообразия. При этом даже довольно большие узкие места могут привести к значительной потере аллельного разнообразия [56]. Возможность исследовать множество локусов позволила оценить эффективный размер популяции с использованием единственной выборки путем оценки степени неравновесного сцепления, вызванного генетическим дрейфом. В 1981 году для этого был предложен метод, но его реализация в естественных популяциях была невозможной до того момента, пока не был разработан способ многолокусного генотипирования [57,58]. Множественное генотипирование микросателлитных локусов позволяет идентифицировать виды на основе их множественных генотипов, и это имеет ряд ценных применений для природоохранной биологии. Например, были разработаны методы для оценки численности популяций, основанные на идентификации видов на основе генотипа множества особей [59]. Огромное количество микросателлитных локусов, присутствующих во многих геномах, позволило создать и картировать сотни микросателлитных локусов у некоторых видов. Это дало начало использованию подходов популяционной геномики в природоохранной биологии. Например, для кукушки *Perisoreus infaustus* было картировано 117 микросателлитных локусов на девять аутосом и Z-хромосому. Первоначальные популяционные исследования с использованием этих маркеров выявили неожиданно высокие уровни нарушения равновесия гамет по всему геному [60]. STR применяются для реконструкции истории миграции и эволюции видов, а также для оценки биологического разнообразия на различных уровнях биологической организации [61].

Микросателлиты имеют две основные особенности. Во-первых, это возможность скрининга многих микросателлитных локусов и определение сильно варьируемых локусов, что позволяет исследовать большое количество полиморфных локусов практически у любых видов. Во-вторых, использование ПЦР позволяет генотипировать микросателлиты практически из любых источников, исключая деструктивные последствия для видов, находящихся на грани исчезновения. Тем не менее, вероятность ошибочной классификации гетерозигот как гомозигот, возникновение нулевых аллелей с последующим затруднением точной оценки полиморфизмов и т.д. заставляет задуматься о необходимости применения этого метода в исследованиях в области природоохранной генетики.

Методы с ограниченной представленностью генома

Для оценки эффективного размера популяции также используют более современный подход популяционной геномики - RADseq [62]. Как было сказано ранее, высокополиморфные микросателлитные маркеры, учитывая, что они доступны для исследуемых видов, обычно могут надежно определять генетическую структуру популяций даже среди близкородственных популяций и предоставлять информацию о генетическом разнообразии популяции, родстве и эффективном размере численности популяции. Однако микросателлитные маркеры имеют ограничения, такие как риск гомоплазии, наличие нулевых аллелей или потенциально недостаточное количество полиморфных локусов у изучаемых видов, что может ограничить их применение. Однонуклеотидные полиморфизмы (SNP), полученные путем секвенирования ДНК, ассоциированного с сайтами рестрикции (RADseq), с образованием тысяч локусов, обеспечивают привлекательную альтернативу, особенно для видов без предварительной генетической информации или для видов и популяций, которые имеют ограниченное количество микросателлитных вариаций. Также, применение RADseq эффективно для больших панмиктических популяций, для которых уровни дифференциации низкие и проведение генетического анализа затруднено. RADSeq превосходит методы с использованием полиморфизмов длины рестрикционных фрагментов (RFLP) и полиморфизмов длины амплифицированных фрагментов (AFLP), которые также используют ограниченное представление генома за счет подвыборки только в сайтах, определенных рестрикционными ферментами. Превосходство заключается в способности проверять и оценивать маркеры одновременно (вместо того, чтобы требовать обширного процесса разработки), и определять, какие маркеры происходят от каждого сайта. RADSeq можно использовать при скрещиваниях любого дизайна и в диких популяциях, что позволяет не только генотипировать и обнаруживать SNP, но и проводить более сложные анализы, такие как количественные генетические и филогеографические исследования [63]. Подход RADseq быстро завоевал популярность, т.к. позволил с низкими затратами обнаруживать и генотипировать тысячи генетических маркеров для любых видов, включая немодельные организмы, что произвело революцию в экологической, эволюционной и природоохранной генетике. А технические характеристики привели к важным соображениям на всех этапах геномных исследований, от конкретных научных вопросов, которые могли быть решены, и затрат на подготовку библиотеки с последующим секвенированием, до типов

систематических ошибок, в прошлом преследовавших другие типы генетических маркеров [35,63].

RADSeq сочетает в себе два метода молекулярной биологии: использование рестрикционных ферментов (рестриктаз) для разрезания ДНК на фрагменты (как и для RFLP и AFLP) и использование молекулярных идентификаторов (MID) для связывания последовательностей определенных биологических видов (рис. 4). ДНК особи разрезается выбранной рестриктазой, образуя набор фрагментов с липкими концами (рис. 4A). Для секвенирования на секвенаторе Illumina эти фрагменты должны быть лигированы с адаптерами, которые будут связываться с проточной ячейкой Illumina. RADSeq использует модифицированные адаптеры Illumina, обеспечивающие связывание и амплификацию фрагментов сайта рестрикции. Фрагменты с липким концом лигируют с адаптером P1, содержащий соответствующий липкий конец и MID. В результате образуется короткая последовательность, которая однозначно идентифицирует определенную особь (рис. 4B). Образцы исследуемых особей баркодируются перед процессом рестрикции и лигирования P1. Помеченные фрагменты рестрикции от ряда особей объединяют, а затем случайным образом разрезают для получения участков со средней длиной в несколько сотен пар оснований (рис. 4C). Вырезанные фрагменты лигируют со вторым адаптером P2 (рис. 4D) и происходит процесс амплификации с помощью ПЦР с использованием праймеров P1 и P2 (рис. 4E). Из-за дивергентной Y-образной структуры адаптера P2, связь с праймером не будет образована, пока не будет завершена амплификация адаптером P1. Затем фрагменты выбираются по размеру (выделяются приблизительно 200-500 основных фрагментов), и полученная библиотека RADSeq впоследствии секвенируется на платформе Illumina. Последовательность генерируется из MID в адаптере P1 и через сайт рестрикционного фермента, генерируя набор данных RAD-тегов (последовательностей ниже сайтов рестрикции), получаемых из значительно уменьшенной части исходного генома. Если сайт рестрикции является симметричным, то из каждого сайта будут созданы два RAD-тега. Платформа Illumina в настоящее время позволяет секвенировать до 150 п.н., и, таким образом, приблизительно 300 п.н., фланкирующих каждый сайт рестрикции, могут быть проверены на полиморфизм.

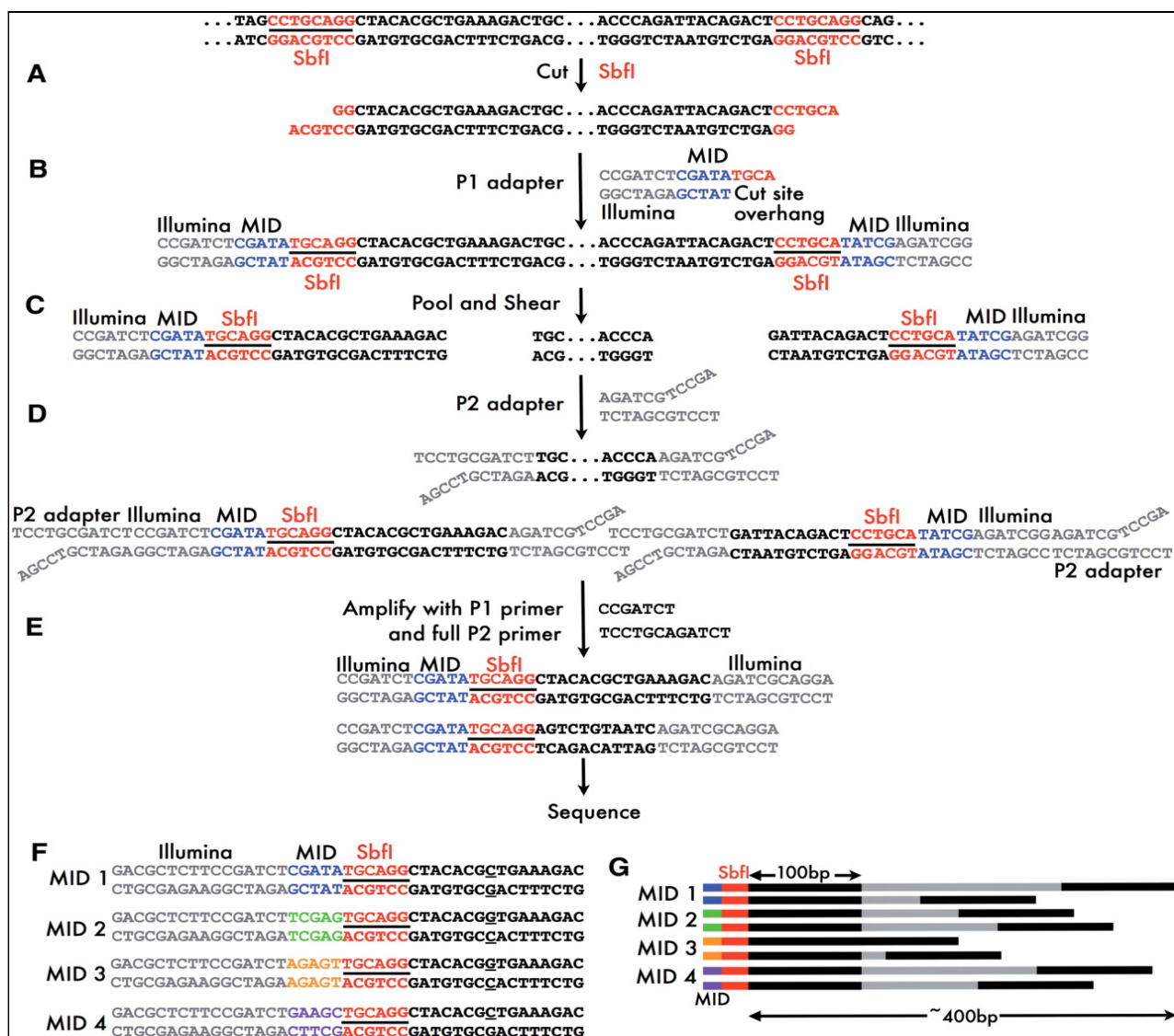


Рисунок 4. Рабочий процесс метода RADseq [63]. (A) ДНК разрезается выбранным ферментом рестрикции. (B) Адаптер P1 лигируется с вырезанными фрагментами. (C) Образцы от нескольких особей объединяются, и все фрагменты случайным образом разрезаются. (D) адаптер P2 лигируется со всеми фрагментами. (E) ПЦР-амплификация с праймерами P1 и P2. (F) Объединенные образцы с разными MID разделяются на SNP. (G) Поскольку фрагменты вырезаются случайным образом, парные концевые последовательности из каждого секвенированного фрагмента будут охватывать область от 300 до 400 п.н. ниже сайта рестрикции.

RADseq активно применяется в различных исследованиях. Например, для проверки потенциала набора данных RADseq для оценки N_e и ее снижения с течением времени, используя методы, основанные на неравновесном сцеплении генов (LD) и частотном спектре аллелей (SFS), было установлено, что данные RADseq эффективны не только для точной

оценки N_e , но и для обнаружения спадов N_e в современных временных масштабах (20 поколений), и что методы, основанные на LD, превосходят их при условии достаточного размера выборки особей [64]. Если доступна референсная геномная сборка с данными о скорости рекомбинации, то методы оценки N_e на основе LD между сцепленными локусами могут быть еще более эффективными [65]. Для проверки генетических эффектов реинтродукции альпийского горного козла *Capra ibex* в Швейцарии, был использован RADseq в генерации более 100000 SNP, что позволило обнаружить заметное уменьшение оценки N_e на основе LD в реинтродуцированных популяциях по сравнению с популяцией источника - близкородственного пиренейского козла *Capra pyrenaica* [66]. Также, было обнаружено, что вывод демографической модели об изменениях N_e на основе данных двойной рестрикции RAD (ddRAD) для двух видов саламандр *Ambystoma talpoideum* и *A. opacum*, согласуется с изменениями размеров популяции, выведенными из данных меток - повторной поимки; поскольку исследование включало секвенирование ddRAD на образцах, собранных несколько десятилетий назад, тенденции времени в N_e можно оценить для этих двух видов, используя как данные о повторной поимке, так и данные о ddRAD [67]. В сравнении вариаций более 2000 SNP в образцах гигантской черепахи *Chelonoidis duncanensis* до и после узкого места, в результате которого N_e сократилось до 150-200 в середине 20-го века, было обнаружено, что масштабы и распределение генетических вариаций в исторических и современных образцах были очень схожими, что объясняется успешной программой ex-situ [68].

Использование молекулярных маркеров для исследований организмов в их естественной среде обитания и оценки воздействия различных антропогенных факторов привело к развитию природоохранной генетики. Тем не менее, многие из таких исследований были ограничены небольшой представленностью генома, что позволяло делать некоторые выводы о количестве генетических вариаций, но затрудняло обобщение данных о видах и их демографической истории. Полногеномные наборы данных повышают эффективность традиционных анализов, которые используют небольшое количество нейтральных локусов для оценки структуры популяции, и позволяют исследовать небольшую долю локусов в геноме, которые демонстрируют экологически значимую адаптацию [69]. Огромные успехи в технологии секвенирования следующего поколения (NGS) привели к возможности обстоятельного анализа полногеномных данных в оценке генетической изменчивости [69–71]. При изучении генетических аспектов популяций немодельных биологических видов, желательно использовать как можно больше образцов особей для получения достоверных результатов. Но секвенирование большого количества образцов остается непомерно

дорогостоящим, особенно если последовательность генома ранее была неизвестна. Кроме того, для целей многих исследований, получение полногеномных данных всех особей весьма избыточно, т.к. это лишь увеличивает вычислительные и биоинформатические затраты. Важным недавним достижением стала разработка подходов к генотипированию путем секвенирования (GBS), которые позволяют секвенировать лишь необходимую часть генома с помощью технологии NGS. Подмножество генома, которое необходимо секвенировать используя подходы GBS, может быть нацелено на использование рестриктаз, зондов гибридизации или на секвенирование транскриптома [72]. По мере дальнейшего развития технологий секвенирования, вычислительных и биоинформатических методов, полногеномное ресеквенирование может стать преобладающим методом экологической и природоохранной геномики. В настоящее время подходы с ограниченным представлением предлагают возможность не только обнаруживать генетические варианты особей, такие как SNP, но также и генотипировать их.

Методы генотипирования путем секвенирования предлагают большие преимущества для популяционной геномики, проводя полногеномный скрининг полиморфизмов, ассоциированных с отбором, рекомбинацией, мутациями и вариациями в дрейфе генов, которые приводят к различным последствиям генетической изменчивости. Как было сказано ранее, большинство исследований в области экологической и природоохранной генетики опирались на небольшое количество предположительно нейтральных молекулярных маркеров (например, аллозимов, микросателлитов и др.), охватывающих очень ограниченное подмножество генома. Эти наборы данных можно использовать для изучения демографических факторов, влияющих на весь геном (например, генетическое разнообразие, поток и дрейф генов, эффективные размеры популяций), но основная проблема заключалась в ограниченной способности исследовать конкретные локусы, которые подверглись отбору и адаптивной эволюции. GBS позволяет идентифицировать не только конкретные области генома, которые, возможно, подверглись естественному отбору, но и повышает точность демографических выводов, за счет значительного увеличения количества анализируемых, предположительно нейтральных, маркеров. Некоторые нейтральные маркеры не могут идентифицировать отдельные популяции, которые в ходе эволюции приобрели устойчивость к определенным патогенам или адаптированы к месту их обитания [73,74]. С другой стороны, нейтральные маркеры могут определить значительную дифференциацию популяции на основе ограниченного потока или дрейфа генов, но селективные области генома могут указывать на адаптивное сходство, сохранившиеся после изоляции или развившиеся параллельно следующей адаптации к новой среде обитания [75,76].

Ряд исследований используют полногеномный анализ для поиска потенциально адаптивных генетических вариаций в контексте популяционной геномики, а также для оценки различных демографических параметров. Рассматриваются виды растений, морских беспозвоночных, морских и пресноводных рыб и мелких млекопитающих, которые дают новые выводы относительно отбора в естественных популяциях в дополнение к измерению демографической истории с использованием нейтральных маркеров [77–80]. Эффективное использование GBS показано в исследовании для филогенетической реконструкции видов цихлид двух родов, *Pundamilia* и *Mbipia* [81]. GBS используется для идентификации участков генома, участвующих в гибридизации, видообразовании и дивергентной адаптации [81,82]. Такие исследования популяционной геномики сталкиваются с проблемами при поиске компромиссов в секвенировании либо отдельных локусов, либо одной особи или нескольких представителей одной популяции. Выбор секвенирования в GBS между глубиной охвата на локус и большими размерами исследуемой выборки, для наиболее эффективного использования данных для популяционной генетики, весьма неоднозначен, требуется индивидуальный подход выбора к определенным исследованиям [83]. Получение геномных данных давно стало необходимостью и подходы NGS значительно улучшили поиск SNP для многих немодельных организмов [84]. В частности, GBS стал высоконадежным подходом для идентификации SNP как внутри, так и между популяциями [85]. GBS эффективен для диплоидных видов, хотя были описаны успешные попытки идентифицировать SNP у четырех тетраплоидных видов [86]. Тем не менее, остаются проблемы для определения генотипов SNP для отдельных полиплоидных организмов, поскольку они могут иметь несколько копий разных аллелей.

Полногеномный подход

Достижения, полученные с помощью технологии высокопроизводительного секвенирования (HTS), выглядят перспективными для исследований, основанных на данных природоохранной геномики. Тем не менее, отмечаются ряд недостатков. Например, короткие последовательности длиной ~100 п.н., получаемые с помощью современных технологий секвенирования, являются проблематичными для сборки генома и обнаружения генетических вариантов. Относительно высокая частота ошибок существующих платформ секвенирования требует получения высокой глубины охвата для правильной идентификации генетических вариантов, а стоимость секвенирования по-прежнему остается дорогой для популяционных исследований в природоохранной биологии, требующих анализа нескольких особей [87]. В настоящее время существует несколько альтернатив для преодоления ограничения стоимости. Первой альтернативой является использование методов секвенирования с ограниченной представленностью генома (RRS), использующие лишь часть генома [88]. Второй альтернативой служит получение данных полногеномного ресеквенирования (WGR) из объединенной эквимольной немеченой ДНК нескольких особей (Pool-seq), или получение данных WGR отдельных особей из популяции с низким охватом (lcWGR) (рис. 5) [89]. Подходы Pool-seq и lcWGR требуют наличия референсного генома. Это, наряду с по-прежнему высокой стоимостью секвенирования и большим спросом на вычислительные ресурсы и хранение данных, ограничивает их применение для немодельных видов в природоохранной биологии [2]. Данные подходы позволили успешно провести скрининг нескольких локусов генома у некоторых видов и сыграли важную роль в решении множества вопросов молекулярной экологии [75]. Однако эти методы имеют свои собственные ограничения и источники ошибок, которые следует минимизировать для верной интерпретации получаемых данных [2].

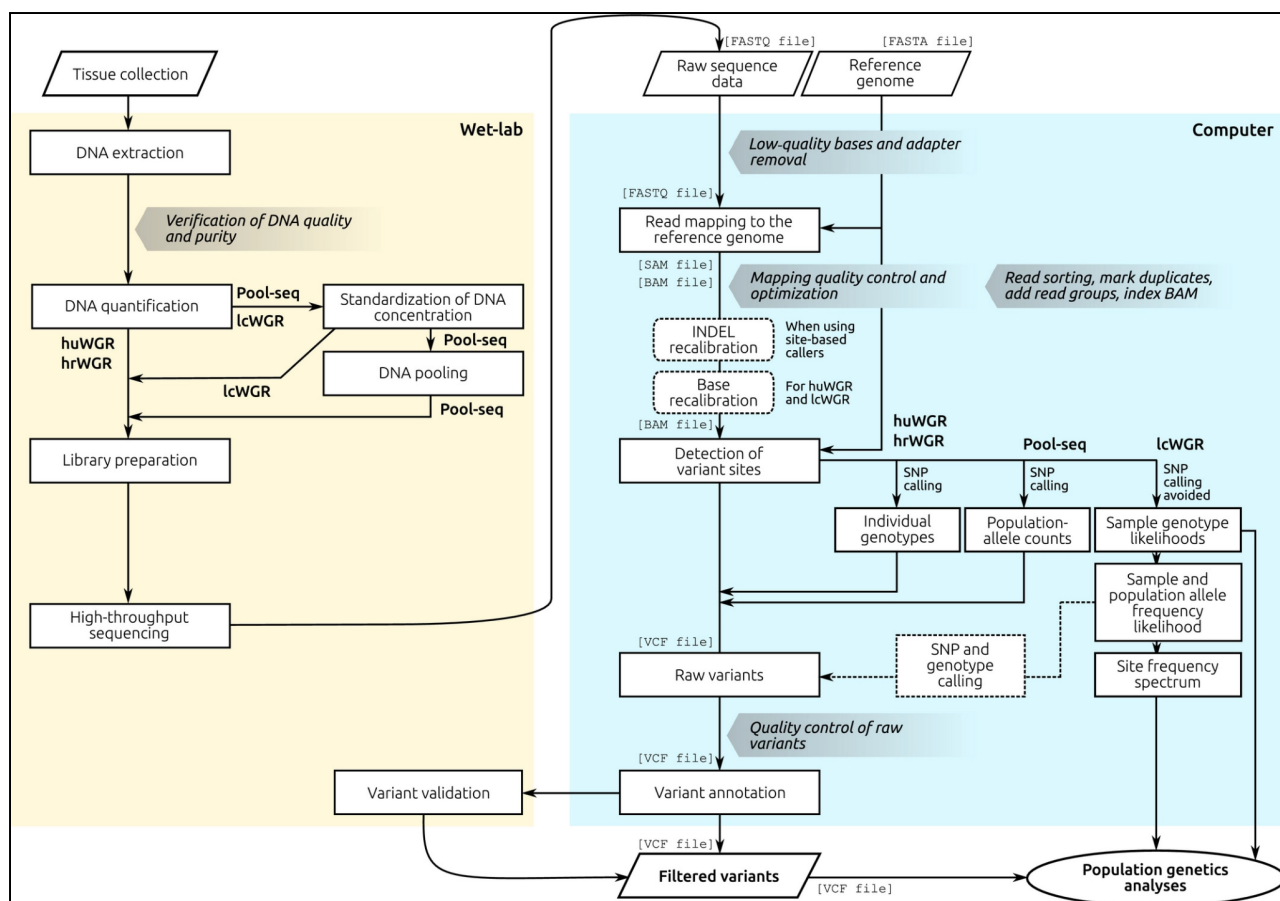


Рисунок 5. Общая схема получения и анализа данных WGR [2]. Показаны четыре подхода к WGR: секвенирование индивидиумов с высокой степенью охвата либо неразрешенными (huWGR), либо разрешенными гаплотипами (hrWGR), секвенирование популяционных геномов на большую глубину путем объединения эквимольных количеств немеченой ДНК (Pool-seq) и секвенирование нескольких особей из популяции на малую глубину (lcWGR). Пунктирные линии обозначают необязательные шаги, заштрихованные стрелки указывают шаги точки проверки качества.

Исследование в области популяционной геномики может быть основано на анализе данных отдельных последовательностей (индивидуальные подходы) или на анализе последовательностей нескольких особей в целом (популяционные подходы). В отдельных подходах, цель заключается в получении высококачественных индивидуальных генотипов, необходимых для реконструкции демографической истории, оценки эффективной численности популяции, оценки генетического разнообразия и др. В настоящее время существует два метода: индивидуальное WGR с высоким охватом и неразрешенными гаплотипами (huWGR) и индивидуальные WGR с высоким охватом и разрешенными гаплотипами (hrWGR). В обоих методах высокая глубина считывания (~30-50x) нацелена на

получение SNP, инсерций/делетий (Indels) и определение генотипа, поскольку высокая глубина считывания помогает отличить реальную генетическую вариацию от ошибки секвенирования [90]. В методе huWGR короткое считывание данных для каждой особи приводит к получению нефазированных индивидуальных генотипов, используемые для последующего анализа. Если требуется информация о гаплотипах, фазирование может быть достигнуто с использованием статистических методов, основанных на генотипах нескольких связанных или не связанных между собой особей. Такой метод ограничивается необходимостью больших размеров выборки и неравновесным сцеплением, варьирующиеся в зависимости от генома. В hrWGR цель состоит в том, чтобы напрямую получить разрешенные по гаплотипам геномы отдельных особей с использованием определенных экспериментальных подходов и короткого и/или длинного считывания, подразумевая, что большой размер выборки не требуется [91].

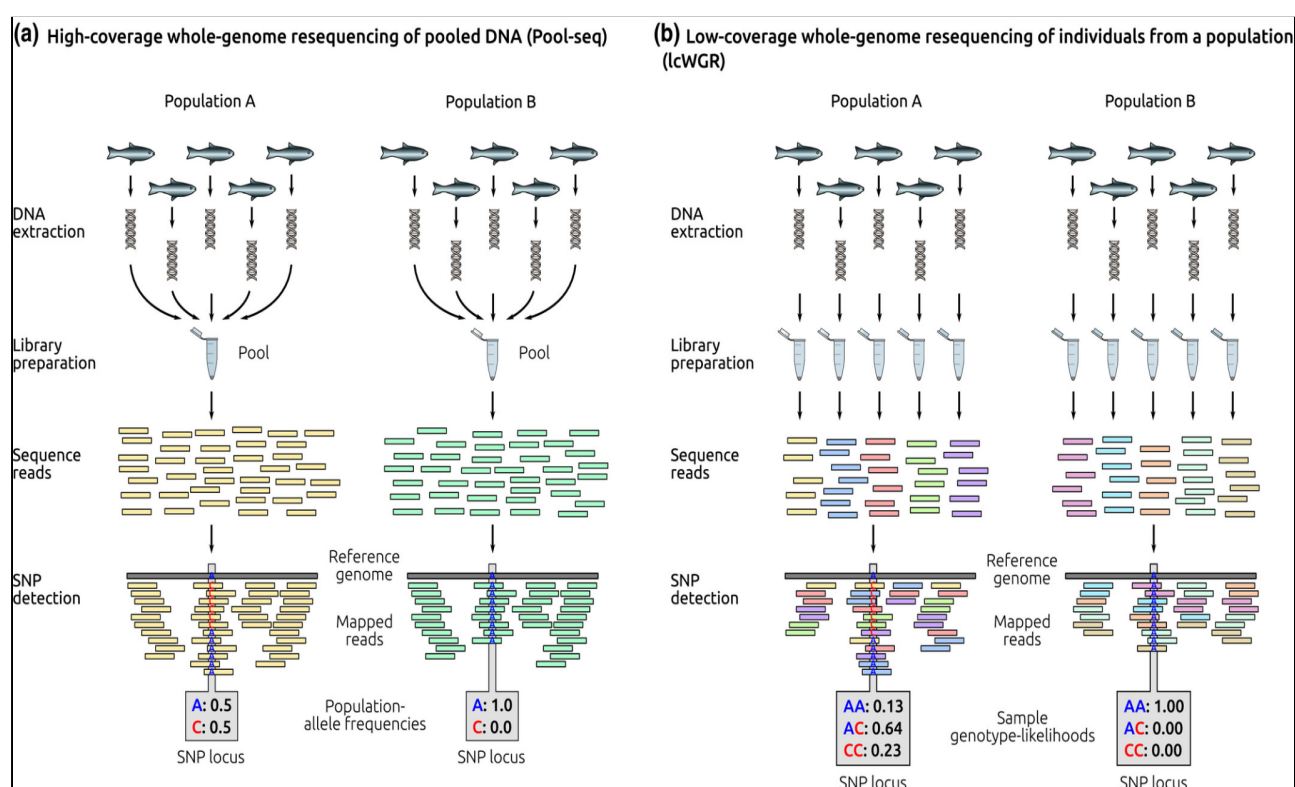


Рисунок 6. Сбор данных в современных популяционных методах WGR. (a) Pool-seq начинается с объединения в одной пробирке эквимольного количества ДНК нескольких особей из популяции. Полученные библиотеки готовятся к секвенированию и каждой популяции присваивается один баркод. Библиотека объединенной ДНК секвенируется с высокой степенью охвата ($> 50x$). Обнаружение и оценка генетических вариантов требует выравнивание прочтений с референсным геномом. Различия в частотах аллелей между

популяциями определяется по числу аллелей для данного полиморфного локуса. (b) *lcWGR* начинается с подготовки отдельной библиотеки для каждой баркодированной особи. Отдельные библиотеки ДНК секвенированы с низкой глубиной охвата (~1-4x). Выравнивание прочтений с референсным геномом требуется для обнаружения генетических вариантов и определения вероятности генотипа выборки (*GL*), основанные на аллелях, присутствующих в отдельных прочтениях.

Pool-seq и *lcWGR* имеют несколько преимуществ и недостатков. Основным преимуществом *Pool-seq* является снижение затрат, достигаемое за счет подготовки одной библиотеки секвенирования на объединенную ДНК нескольких особей, вместо одной библиотеки на каждую особь. Это позволяет использовать большие размеры выборки (рис. 6a). Кроме того, объединение равных количеств ДНК нескольких особей способствует многократному секвенированию нескольких хромосом, что приводит к улучшению оценок генетического разнообразия [89]. В *Pool-seq*, обнаружение и оценка генетических вариантов является производной от относительной доли считанных аллелей в пределах одного пула (рис. 6a). *Pool-seq* имеет три основных ограничения: во-первых, отдельные генотипы не учитываются после смешивания образцов ДНК в пуле. Это делает невозможным отслеживание технических ошибок во время подготовки библиотеки. Во-вторых, оценка частоты аллелей зависит от множества факторов, включая неравномерное представление отдельных ДНК в пуле, а также от ошибок секвенирования. Наконец, редкие аллели, вероятно, будут недостаточно представлены в наборах данных такого типа, что может привести к ограниченному распределению частот аллелей или частотного спектра сайта (SFS). Таким образом, *Pool-seq* эффективно выявляет высокочастотные аллели, и плохо - низкочастотные, соответственно [92]. Получение полного SFS имеет важное значение в популяционной генетике, так как этот показатель отражает все генетические вариации в несвязанных сайтах. Его форма варьируется в зависимости от различных эволюционных процессов, включая узкие места и естественный отбор, а также используется для вывода нескольких показателей, включая *D* и *Fst* Таджимы [93–95]. Было предложено несколько способов смягчить имеющиеся ограничения. Знание структуры популяции может помочь избежать случайного объединения особей разного происхождения. Однако смешанные агрегации также могут быть косвенно обнаружены из данных *Pool-seq* и при необходимости исключены из апостериорного анализа, поскольку они демонстрируют большую длину ветвей в филогенетическом дереве на основе попарных генетических дистанций [96]. В отсутствие предварительной информации о структуре популяции, экологические и

биологические знания (например, средний возраст поколения, стадия зрелости на момент отбора проб, размер и возраст) могут помочь с выводом о локальном происхождении особей. Вышеуказанные ограничения, касающиеся отсутствия гаплотипа и, следовательно, информации об отдельных генотипах и неравновесном сцеплении, могут быть преодолены путем выбора подмножества информативных SNP, идентифицированных из данных Pool-seq с последующим генотипированием ряда особей в популяции по этим информативным SNP с использованием методов генотипирования на основе ПЦР или NGS. Данный этап проверки SNP рекомендуется выполнять как часть протокола надлежащей практики для рабочего процесса Pool-seq [2].

Основным преимуществом lcWGR является низкая глубина секвенирования (~1-4x), что облегчает анализ большого количества особей из одной популяции, но, тем не менее, для каждой особи необходимо подготавливать одну библиотеку (рис. 6b). Несмотря на значительную экономию средств с помощью этой процедуры, lcWGR по-прежнему остается дороже, чем Pool-seq для использования эквивалентного размера выборки и глубины секвенирования [2]. Как и в случае с Pool-seq, некоторые из недостатков lcWGR заключаются в том, что отдельные генотипы не могут быть определены: низкая глубина для каждой особи препятствует надежному определению генетических вариантов и генотипа. lcWGR выявляет вариантные сайты и вычисляет правдоподобие генотипа (GL) для каждого сайта на основе данных последовательностей нескольких особей в образце с использованием вероятностной структуры [97].

Подходы huWGR и lcWGR используются для реконструкции демографической истории различных видов. Так, были проанализированы данные huWGR 34 панд (*Ailuropoda melanoleuca*) и обнаружены генетические аспекты множества демографических событий, таких как увеличение популяции, узкие места и расхождения [98]. Было определено, что различные антропогенные факторы, скорее всего, способствовали их сокращению за последние ~3.000 лет. Реконструкция демографической истории пяти различных экотипов косаток (*Orcinus orca*) на основе данных lcWGR 48 особей и данных huWGR 2 особей помогла понять, что модели дифференциации между парами современных аллопатрических и симпатрических экотипов, скорее всего, являются следствием экологического расхождения и генетического дрейфа в результате узких мест, произошедших во время прошлых событий [99]. С помощью данных huWGR возможно оценить значение N_e . Например, с использованием модели парного последовательного марковского слияния (PSMC) были проанализированы оценки N_e четырех видов мухоловки из рода *Ficedula* на основе полногеномных данных 200 особей из 10 европейских популяций [100]. Было определено,

что самый поздний общий предок четырех видов датируется 1-2 млн лет назад, и каждый вид следовал отдельным эволюционным путем, включающим рост популяции, сокращение (~100.000-200.000 лет назад) и экспансию. Предположено, что среднее покрытие генома $\geq 18\times$ на особь, порог фильтрации на уровне ≥ 10 чтений и не более 25% отсутствующих данных необходимы для правильного вывода демографической истории с использованием данных PSMC и huWGR [100]. Более поздние статистические модели обещают возможность оценить N_e по данным lcWGR и Pool-seq, одни из которых основаны на инбридинге по идентичности происхождения (IBD), а другие на изменениях частоты аллелей между двумя временными выборками с поправкой на потенциальное увеличение дисперсии в частоте аллелей из-за двух этапов выборки, задействованных в экспериментах Pool-Seq (т. е. во время выборки индивидуумов и секвенирования пулов) [101,102]. Таким образом, приведенные выше исследования показывают, как huWGR и lcWGR могут облегчить анализ динамики N_e и отслеживание исторических демографических событий, которые могут помочь понять закономерности генетического разнообразия и структуры в современных популяциях. Эти виды анализа можно распространить на немодельные виды, представляющие интерес для природоохранной биологии. Геномы с определением гаплотипов, полученные с помощью подходов hrWGR, обещают повышенную точность демографической истории и оценки эффективного размера популяции [2].

Реконструкция демографической истории

Изучение демографической истории вида, включая бутылочные горлышки численности, различные модели миграции, расширение ареала и изменения исторической эффективной численности популяции (N_e), представляет большой интерес с точки зрения природоохранной биологии, поскольку помогает понять ранние демографические события и их влияние на генетическое разнообразие современных популяций. Такие исследования позволяют получить данные о влиянии датированных экологических событий (например, снижение потока генов, влияния антропогенных факторов, изменения климата и др.) на процессы демографической истории. Понимание тенденций численности популяций является одним из ключевых в исследованиях природоохранной биологии. Изучая демографическую историю, можно сделать вывод о разнообразных эволюционных процессах, включая расселение, видообразование и сокращение популяции, которое может привести к риску исчезновения [69]. Для возможности реконструкции демографической истории были достигнуты значительные успехи в разработке соответствующих методов популяционной геномики, основанных на анализе последовательностей генома из небольших выборок особей и имеющих общую теоретическую основу: парное последовательное марковское слияние (PSMC) и множественное последовательное марковское слияние (MSMC) [103,104]. Метод PSMC может использовать для анализа данных нефазированных последовательностей от одной особи, тогда как MSMC может использовать последовательности от нескольких особей. Эти два метода особенно полезны при изучении более ранних временных масштабов и при очень ограниченном количестве выборок [105]. Так, PSMC был эффективно использован при изучении отдельных древних образцов, в том числе образцов древней лошади рода *Equus*, древнего волка *Canis dirus/Aenocyon dirus* и двух шерстистых мамонтов *Mammuthus primigenius* [106–108].

Помимо реконструкции демографической истории, PSMC и MSMC могут использоваться для определения времени расхождения популяций и оценки скорости мутаций по древним геномам. Однако различные исследования показали, что результаты из двух методов могут быть чувствительны к интерпретации результатов, т.к. методы не предлагают явной основы для проверки гипотез [109]. Как следствие, необходимо некоторое понимание основ биологической теории и статистических методов, чтобы гарантировать достоверность результатов интерпретации.

Методы и их ограничения

Методы PSMC и MSMC изначально были разработаны для получения информации о ранних изменениях эффективного размера популяций, включая время расхождения и определение узких мест. Результатом каждого метода является график, показывающий, как эффективный размер численности популяции, к которому принадлежит исследуемая особь, изменился с течением времени. PSMC и MSMC могут комбинировать информацию из гораздо большего количества локусов, чем это возможно с вычислительной точки зрения с помощью древовидных методов, таких как Skyline-plot [110]. По этой причине методы последовательного марковского слияния могут исследовать более глубокие временные рамки, потому что данные с гораздо большей вероятностью будут включать локусы, у которых есть самый последний общий предок из далекого прошлого. Например, применение PSMC позволило пролить свет на более чем миллион лет демографической истории современных людей и других человекообразных обезьян [111]. Хотя ни PSMC, ни MSMC не предоставляют формальную основу для проверки конкретных гипотез о причинах изменений размера популяции. MSMC может использоваться вместо PSMC, даже когда для анализа доступен только один диплоидный геном. MSMC верно оценивает скорость рекомбинации, когда PSMC не может этого сделать [103]. Если доступно несколько геномов, MSMC имеет больше возможностей для определения недавних изменений в эффективном размере популяции, потому что добавление аллелей увеличивает вероятность того, что в недавнем прошлом произойдет событие слияния. При анализе геномов человека MSMC информативен для событий, которые произошли совсем недавно, ~2.000 лет назад, тогда как PSMC не дает надежного определения изменений в эффективном размере популяции, которые произошли за последние ~20.000 лет [103].

Существуют ограничения, насколько надежные оценки демографической истории могут дать модели последовательного марковского слияния. Чтобы модели могли предоставить результат о размере популяции в заданный период времени, в анализируемом геноме должны быть локусы, в которых аллели сливаются в этот заданный период. Однако теория слияния показывает, что аллели с глубокими слияниями относительно редки [109]. Когда анализируется демографическая история, маловероятно, что заранее известны какие-либо данные о скорости слияния. Следовательно, вывод о темпах слияния и, следовательно, о размерах популяции становится гораздо более шумным для ранней демографической истории. Важным соображением при использовании PSMC и MSMC

является то, что их результаты не всегда могут быть надежно интерпретированы как графики изменений размера популяции. Модели оценивают скорость слияния в каждый момент времени. Теория коалесценции (Coalescent theory) показывает, что величина, обратная этой скорости, иногда называемая обратной мгновенной скоростью слияния, может использоваться в качестве показателя размера популяции при определенных допущениях [112]. Тем не менее, если исследуемая популяция не соответствует этим предположениям, то на скорость слияния могут влиять другие факторы, и в процессе интерпретации очевидных изменений в размере популяции их необходимо учитывать [112]. Например, связь между временем слияния и размерами популяции может быть нарушена естественным отбором и неслучайным спариванием. Одним из наглядных примеров неслучайного спаривания является модель *n*-островков, где *n* панмиктических популяций разделены, за исключением некоторой фиксированной скорости миграции. Поскольку два аллеля не могут слиться, пока они находятся на разных островах, ожидаемое время слияния определяется количеством островов и скоростью миграции между ними, а также размером популяции [112]. В частности, когда скорость миграции между островами низка, сливающийся эффективный размер популяции - параметр, определяемый последовательными методами, - может быть намного больше, чем истинный размер популяции [103]. Таким образом, пики на демографическом графике могут соответствовать периодам увеличения структуры популяции, а не увеличению численности.

Предположение о панмиксии также может быть нарушено инбридингом, который увеличивает скорость слияния и, следовательно, снижает эффективный размер популяции. Инбридинг проявляется в геноме в виде длинных участков гомозиготной последовательности (рис. 14b: протяженные гомозиготные участки с низким генетическим разнообразием в аутосомах образца *M. nivalis* из Дальневосточной популяции, свидетельствующие о недавнем инбридинге). Одна из стратегий проверки того, повлиял ли инбридинг на результат реконструкции демографической истории, состоит в том, чтобы идентифицировать и удалять такие серии гомозиготности, затем повторять анализ и проверять любые изменения в результатах [113]. В целом, очень трудно провести различие между изменениями, которые связаны с изменением численности популяции, и изменениями, которые вызваны другими демографическими событиями (рост миграции, изменение структуры популяции и др.). Изменения некоторых демографических событий не привязанных к временным точкам, могут изменить кривую демографической истории. Так, демографическая история показывает подъемы и спады в течение многих тысяч поколений из-за изменения скорости миграции изменяется в постоянном размера *n*-островков популяции [112]. Когда за короткий

промежуток времени происходит много изменений, они могут взаимодействовать сложным образом, поэтому сложно связать какую-либо особенность полученных результатов демографической истории с конкретным изменением. Применение PSMC и MSMC для высокоструктурированных популяций может вызвать затруднение в процессе интерпретации полученных результатов и их достоверность может быть сомнительна [109]. Влияние структуры популяции нельзя оценить непосредственно по наблюдаемому времени слияния. Даже если размер популяции постоянен, но время слияния меняется исключительно в результате изменений в структуре популяции, есть вероятность нахождения ложных изменений размера популяции. Кроме того, для тестирования структуры популяции обычно требуются данные от многих особей, которые могут быть недоступны при использовании методов, основанных на последовательном марковском слиянии [112].

На данный момент разработаны новые методы, расширяющие структуру, обеспечиваемую последовательным марковским слиянием [105]. Эти методы могут в некоторых аспектах превзойти PSMC и MSMC. Метод последовательного марковского слияния с добавлением множества нефазированных образцов (SMC++) позволяет использовать гораздо большее количество геномов, чем это возможно для MSMC, может использовать нефазированные данные и не требует поэтапного разделения геномов [114]. SMC++ более точен, чем MSMC, особенно для определения размеров популяции в недавнем прошлом. Метод геномного вывода о слиянии с минимальным допущением (MAGIC), аналогичен последовательному марковскому слиянию [115]. Однако MAGIC не использует явную модель слияния и рекомбинации и может оценивать множество различных параметров на основе эмпирических данных. Учитывая разные сильные и слабые стороны различных методов реконструкции демографической истории, необходимо использование нескольких методов с последующим сравнением полученных результатов [105].

Важной задачей популяционной генетики является определение времени расхождения между близкородственными видами или популяциями. Один из подходов, который можно использовать для решения данной задачи, - это объединение нескольких видов [116]. Однако для этого подхода требуются данные от нескольких особей каждого вида, и полученный набор данных будет содержать большое количество локусов, что осложняет вычисления. Методы PSMC и MSMC можно использовать для определения времени расхождения при включении информации из полногеномных данных. Хотя это не является предполагаемым использованием методов, PSMC можно адаптировать для определения точки, в которой прекращается поток генов между парой популяций [103]. Простой подход - сравнить графики PSMC, полученные от особей двух видов или популяций. Точка, в которой их участки

становятся идентичными, указывает, когда они представляют одну и ту же предковую популяцию (рис. 7а; рис. 18 : ~600.000 лет назад Новосибирская и Дальневосточная популяции составляли одну и ту же предковую популяцию на протяжении ~300.000 лет).

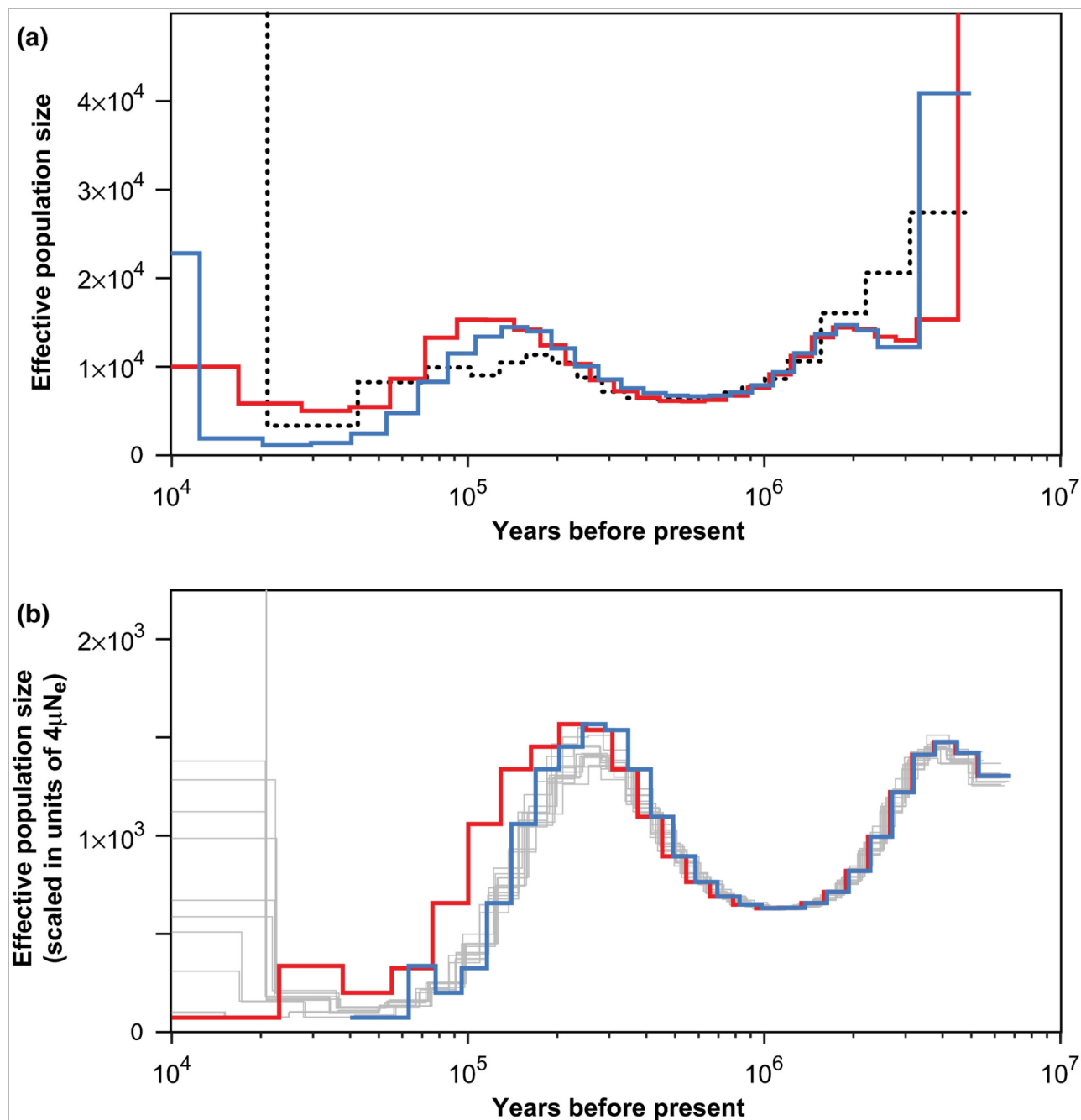


Рисунок 7. Визуализация двух различных вариантов использования методов PSMC и MSMC [109]. (а) Датировка событий видообразования с использованием PSMC [103]. Сплошные линии представляют собой графики PSMC для геномов двух современных людей: йоруба (красный) и китайца (синий). Оглядываясь назад во времени, графики становятся идентичными ~100.000-120.000 лет назад, что указывает на общую демографическую

историю. Пунктирная линия показывает график PSMC гибридного генома, построенного из X-хромосом геномов йоруба и китайца, в масштабе 0.75. Если заглянуть в будущее, то бесконечный размер популяции ~20.000 лет назад предполагает прекращение потока генов между двумя популяциями. (b) Оценка частоты мутаций с помощью PSMC [117]. Две выделенные линии представляют собой графики для современного человека возрастом 45.000 лет из Усть-Ишима в Западной Сибири, либо нескорректированные (красный), либо сдвинутые по горизонтали (синий) для сопоставления с графиками современных неафриканских людей (серые линии). Величина горизонтального смещения указывает на количество мутаций, произошедших за 45.000 лет, что дает возможность оценить скорость мутаций.

Альтернативный подход на основе PSMC использует синтетический диплоидный геном, который состоит из фазированных или нефазированных гаплотипов из двух видов или популяций. Синтетический геном имитирует геном гибрида F1, поэтому этот подход известен как hPSMC [118]. Слияние двух аллелей в каждом локусе может происходить только в предковой популяции. Поскольку скорость слияния падает до нуля после того, как популяции становятся изолированными друг от друга и зон симпатрии не возникает, эффективный размер популяции будет бесконечен (рис. 7a). Этот метод может обеспечить только максимальный предел времени расхождения, потому что возможно, что популяции разошлись позже, чем самое последнее событие слияния. С другой стороны, сроки расхождения могут быть ошибочными, если популяции не достигли полной репродуктивной изоляции. Учитывая чувствительность hPSMC к любому потоку генов, который произошел после расхождения между двумя популяциями, этот метод целесообразно использовать для определения даты прекращения потока генов, а не расхождения популяции [118]. Множественное последовательное марковское слияние лучше работает в анализе различных паттернов расхождения популяций и может использоваться для получения более полной картины того, как поток генов между популяциями менялся с течением времени. MSMC может рассчитать скорость перекрестного слияния между каждой парой популяций, а также скорость слияния внутри каждой популяции, если известно, какие образцы принадлежат конкретной популяции [119]. Соотношение этих показателей будет увеличиваться или уменьшаться в зависимости от потока генов между субпопуляциями: если миграция между популяциями высока в течение некоторого периода времени, коэффициент перекрестного слияния между популяциями должен постепенно увеличиваться в течение этого интервала, поскольку популяции будут иметь больше общих аллелей [104].

Метод PSMC также использовался для оценки скорости мутаций по полногеномным данным [117]. Для этой задачи требуются две последовательности, которые имеют разный возраст и принадлежат одной и той же популяции. Учитывая их общую демографическую историю, два генома должны давать очень похожие результаты оценки эффективного размера популяции. Тем не менее, два участка будут смещены вдоль оси времени (X) (рис. 7b). Скорость мутаций оценивается количеством найденных мутаций, которые впоследствии добавляются к древним последовательностям для сопоставления двух графиков, затем полученное число делится на разницу в возрасте двух образцов. Это можно сделать численно, добавляя мутации в более старый геном до тех пор, пока его демографическая история не станет идентичной с демографической историей более молодого генома, но оценку можно провести более строго, используя подход максимального правдоподобия (ML) [117]. Определение скорости мутаций с использованием PSMC необходимо производить только в том случае, если оценка скорости мутаций для исследуемого вида в противном случае недоступна. Использование графиков демографической истории для этой цели зависит от множества сбивающих с толку факторов, включая те, которые влияют на стандартное использование PSMC для определения изменений размера популяции. Другие, более общепринятые методы оценки скорости мутаций, будут значительно более точными [120].

Упомянутые методы предоставляют мощные средства реконструкции демографических историй на основе геномных данных. Они особенно полезны, когда последовательности генома ограничены только несколькими особями или когда цель состоит в том, чтобы исследовать временные рамки, недоступные для других методов [121]. Однако при интерпретации результатов PSMC и MSMC возникают серьезные проблемы, поскольку для проверки выдвинутых гипотез из полученных результатов, обычно требуется больше данных, чем доступно при использовании этих методов. Тем не менее, зная теоретическую основу и учитывая возможные ограничения, можно получить ценную информацию о демографической истории, которая в ином случае могла бы быть недоступна.

Значимость таких реконструкций

Значимость реконструкции демографической истории для популяционной генетики и в целом для природоохранной биологии сложно переоценить, несмотря на наличие некоторых ограничений в используемых методах. Методы PSMC и MSMC активно используются в научных исследованиях (рис. 8).

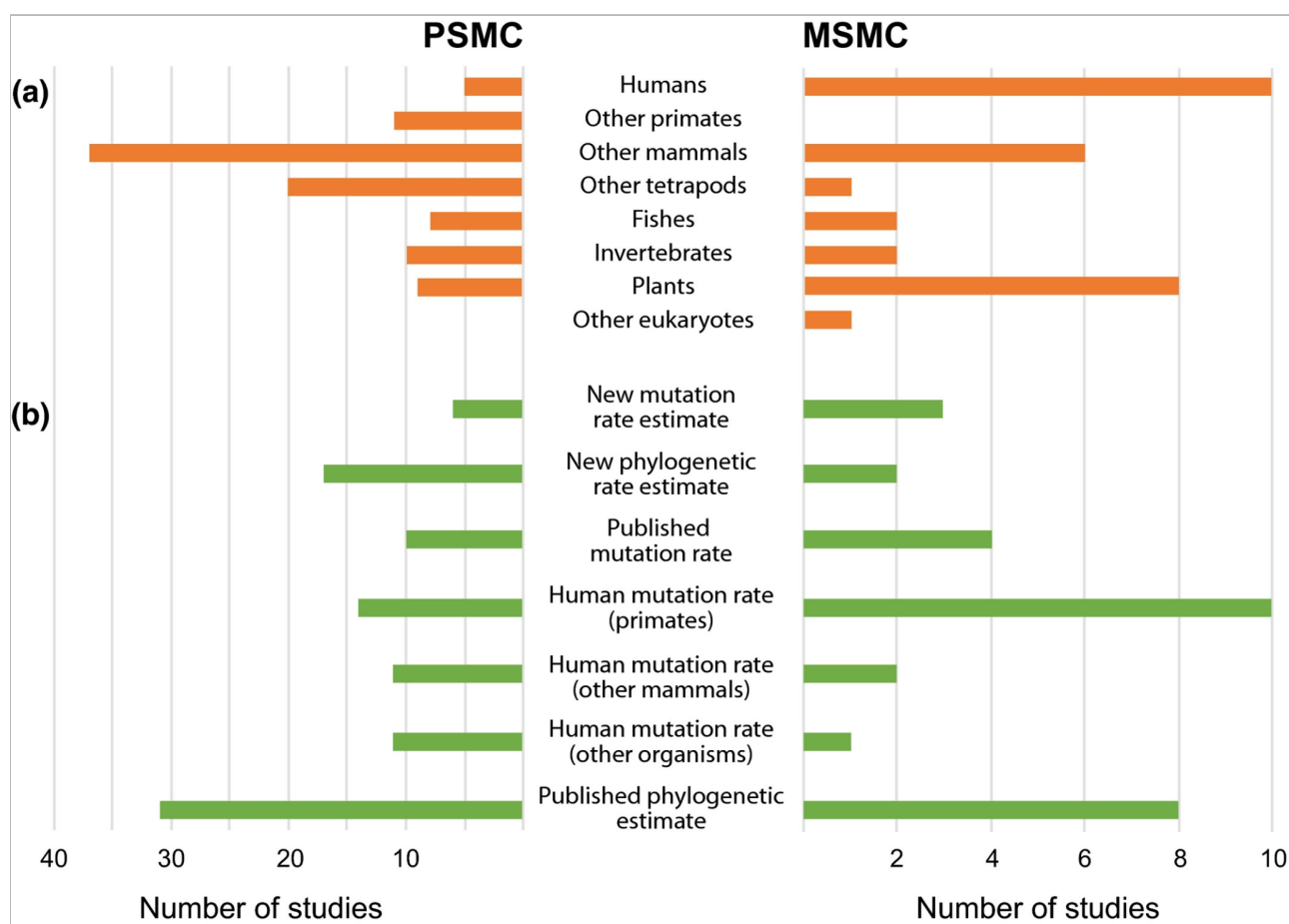


Рисунок 8. Данные случайных выборок из 100 исследований, в которых реализован метод парного последовательного марковского слияния (PSMC), и 30 исследований, в которых реализован метод множественного последовательного марковского слияния (MSMC) [109]. (a) Таксономическая принадлежность изученных организмов. (b) Источник частот мутаций, используемый для масштабирования демографических графиков. Выбор частоты мутаций влияет на масштаб горизонтальной оси и оценку эффективного размера популяции, но не влияет на качественную форму графика. Например, более высокая частота мутаций приведет к линейному уменьшению предполагаемого размера популяции и сдвинет кривую ближе к настоящему времени.

Так, например, был реализован геномный проект по иберийской рыси (*Lynx pardinus*) и проведен анализ демографической истории, эволюции и популяционной генетики рыси на основе полногеномных данных [8]. Иберийская рысь является одной из самых исчезающих видов из семейства кошачьих (*Felidae*) и уникальный объект исследования, находящийся на грани исчезновения. Проведение геномных исследований исчезающих видов имеют важное значение для природоохранной биологии. Благодаря реконструкции демографической истории таких видов и изучению их эволюции, могут быть выявлены закономерности геномной эрозии, ограничивающие жизнеспособность видов, находящихся на грани исчезновения, а также могут быть разработаны инструменты для их эффективного сохранения. В демографической истории иберийской рыси был выявлен ряд серьезных проблем связанных с предковой популяцией, которые предшествовали ее демографическому спаду в 20-м веке, что сильно повлияло на эволюцию ее генома. Одной из таких выявленных проблем является наличие четырех узких мест (рис. 9В), характеризующееся сначала резким падением численности как иберийских, так и близкородственных евразийских рысей (*Lynx lynx*), затем небольшим подъемом с незначительными колебаниями из-за ледникового цикла, что способствовало расширению ареала и образованию зон симпатрии на Пиренейском полуострове, в голоцене на юге Франции и севере Италии, увеличивая поток генов, обнаруженный в анализах расхождения. Последующий период оледенения, вероятно, вюрмского, вызвал второе сокращение популяции, приведшее к изоляции обоих видов рысей. Однако, выявленный генетический обмен не прекращался, о чем свидетельствуют результаты реконструкции демографической истории (рис. 9А). Стоит отметить, что непрекращающийся поток генов после видообразования или расхождения популяций от общей предковой популяции, может наблюдаться не только среди кошачьих, но и других млекопитающих, например, это наблюдается по результатам реконструкции демографической истории *M. nivalis* из данного исследования. Дальнейший анализ реконструкции демографической истории *L. pardinus* позволил зафиксировать третье, более недавнее сокращение, которое привело к сокращению эффективной численности популяции иберийской рыси до критического значения - менее 300 особей около 315 лет назад (рис. 9С). Четвертое и самое последнее узкое место произошло в 20 веке.

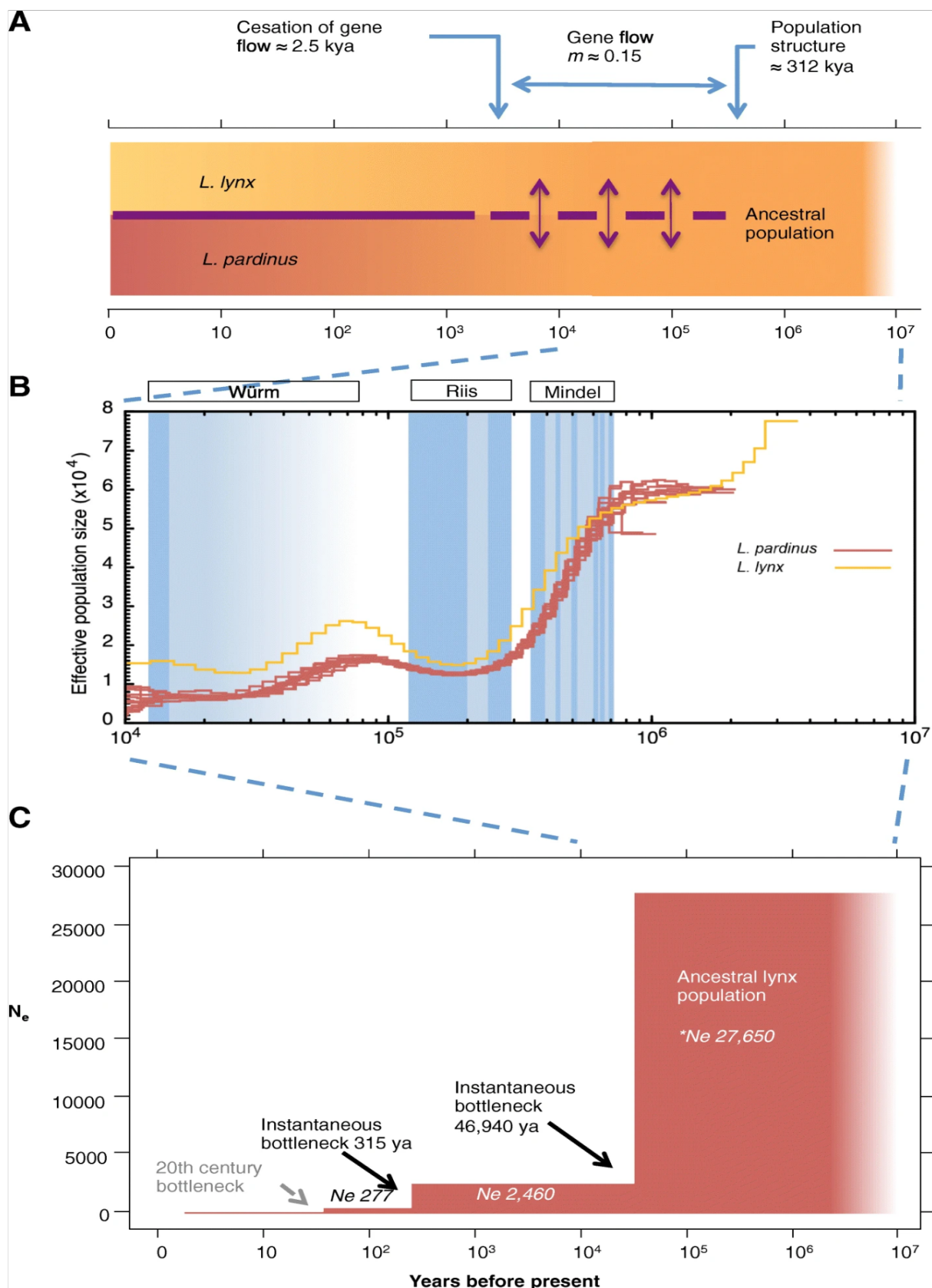


Рисунок 9. Эволюционная и демографическая история иберийской и евразийской рыси [8]. Расхождение, скорректированное в соответствии с моделью расхождения потока генов (А). Результаты свидетельствуют о сравнительно недавней дивергенции двух видов рысей, за

которой последовал период потока генов, продолжавшийся до недавнего времени (~2.000 лет назад). Эффективный размер популяции во времени оценивался для каждой из 11 иберийских рысей и одной евразийской особи с помощью PSMC (B). Демографические истории сходны для десяти иберийских рысей и немного отличаются для евразийской рыси, хотя оба вида следуют в основном параллельным колебаниям численности популяции, вероятно, связанным с ледниковыми циклами. Два узких места (C).

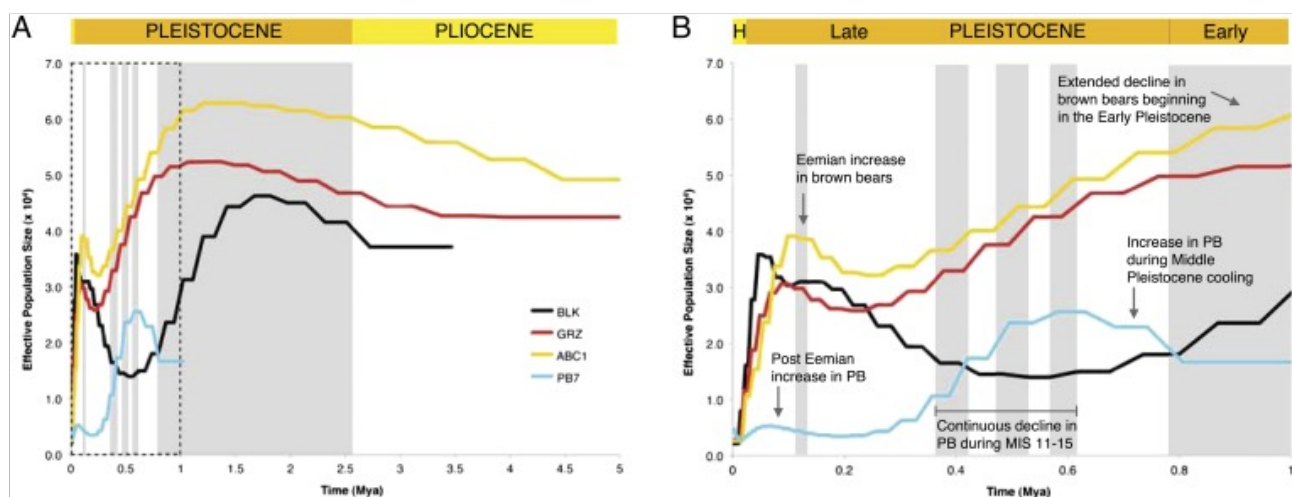


Рисунок 10. Демографическая история медведей [7]. (A) Оценки N_e , сделанные на основе четырех геномов медведей, показанные на временном интервале в 5 миллионов лет. Пунктирная рамка указывает на временной интервал. Большая серая рамка иллюстрирует ранний плейстоцен, тогда как меньшие серые области относятся к ключевым геологическим событиям. (B) Большая заштрихованная область (справа) относится к раннему плейстоцену, а другие серые области (справа налево) - к межледниковым морским изотопным стадиям и эемскому периоду. Стрелки указывают на основные события в истории популяции медведей. H - эпоха голоцена.

Другим примером значимости реконструкции демографической истории служит исследование белых (*Ursus maritimus*) и бурых (*Ursus arctos*) медведей [7]. Белые медведи, хорошо адаптированные к экстремальным условиям Арктики, являются символом угрозы биоразнообразию в контексте глобального изменения климата. Их отличие от бурого медведя представляет собой пример быстрой фенотипической эволюции. Результаты анализа демографической истории медведей (рис. 10), позволили определить их независимую эволюционную историю на протяжении миллионов лет и выявили в эволюционной истории непрерывающийся поток генов между видами медведей, вероятно, образовавшийся в

результате миграции ранее изолированных популяций белого медведя и образовании зон симпатрии, что значительно осложняет задачу определения времени расхождения. Тенденция к потеплению климата в Арктике вызвала резкое отступление морского льда и, судя по результатам демографической истории, ареал белого медведя будет смещаться на сушу, что приведет к образованию гибридов белых и бурых медведей. Гибриды F1 и F2 были зарегистрированы в северной части моря Бофорта в Арктической Канаде, где ареалы белых и бурых медведей, по-видимому, уже перекрываются, возможно, это недавняя реакция на климатические изменения. Подобное скрещивание с другими арктическими видами может иметь большое влияние на полярное биоразнообразие. Полученные результаты демографической истории также могут свидетельствовать о том, что рефугиум Свальбард, возможно, имел важное значение для прошлых межледниковых периодов, из которого могла произойти реколонизация популяций медведей в последующие более прохладные периоды. Если современные популяции белого медведя являются результатом реколонизации, - это может объяснить выявленное низкое генетическое разнообразие белого медведя, что делает его более уязвимым к будущим климатическим и другим условиям окружающей среды.

Относительно подробное рассмотрение значимости реконструкции демографической истории на примере исследований иберийской рыси (*L. pardinus*), белого (*U. maritimus*) и бурого (*U. arctos*) медведей объясняется тем, что они хорошо показывают необходимость проведения обстоятельного анализа всех биологических видов в контексте природоохранной биологии, для своевременного применения соответствующих мер для сохранения видов и отслеживания тенденции влияния различных биотических, абиотических и антропогенных факторов. Реконструкция демографической истории *L. pardinus* позволил по-новому взглянуть не только на демографические аспекты, но и на эволюционные, выявив недавнее расхождение, продолжающийся поток генов и несколько серьезных проблем с популяцией за последние тысячелетия. Такая демография сформировала паттерны нуклеотидных замен, а генетический дрейф и сопутствующее снижение эффективности естественного отбора привели к чрезвычайно низким уровням генетического разнообразия и высокой генетической нагрузке, что свидетельствует о высоком уровне геномной эрозии иберийской рыси [8]. А реконструкция демографической истории представителей семейства медвежьих (*Ursidae*), позволила выдвинуть некоторые предположения о причинах формирующейся эволюционной тенденции этих видов и наглядно показать влияние климатических изменений на различные биоценозы [7].

Материалы и методы

Использованные в работе данные включают данные секвенирования полногеномных библиотек для 3 образцов *M. nivalis* из условно выбранных популяций (табл. 1), расположенных в более чем 1.000 км друг от друга (рис. 1): сцепленные прочтения Польской популяции (10X); парные прочтения Illumina 150 п.н. для Новосибирской (MNIV) и Дальневосточной (T100) популяций; а также сборку генома обыкновенной ласки на основе сцепленных прочтений 10X Genomics Польской популяции.

Таблица 1. Источники исходных данных.

Данные	Образцы	Сбор образцов и предоставление данных	Выделение ДНК и создание клеточных культур
Сборка на основе сцепленных прочтений 10X Genomics	10X - Польская популяция (Польша, г. Краков)	Jose Melo-Ferreira, Karol Zub Klaus-Peter Koepfli	Inês Miranda, Liliana Farelo
Сцепленные прочтения 10X Genomics			
Парные прочтения Illumina 150 п.н.	MNIV - Новосибирская популяция (Россия, г. Новосибирск)	Виктор Панов, Анна С. Мухачева	Полина Перельман, Наталья Сердюкова, Виолетта Р. Беклемишева
	T100 - Дальневосточная популяция (Россия, Сихотэ-Алинский заповедник)		

Оценка качества и фильтрация данных

Данные сцепленных прочтений, полученные с помощью технологии 10X Genomics сочетают в себе стратегию баркодирования с секвенированием Illumina. Эта технология маркирует все считывания, исходящие из небольшого числа молекул ДНК (от 5 до 10) длиной ~50 тыс п.н. с одним и тем же баркодом. Для качественного проведения анализа, на исходных данных секвенирования образца 10X, полученного на платформе 10X Genomics, была использована программа ЕМА v.0.6.2, с помощью которой было выполнено удаление баркода

из сцепленных прочтений [122]. Оценка качества исходных данных осуществлялась с помощью программ FastQC v.0.11.9 и KrATER [123,124]. FastQC предоставляет подобный отчет в формате HTML, содержащий сводные графики и таблицы для удобной оценки качества прочтений. KrATER является программным инструментом для визуализации покрытия k-меров на основе базы, созданной с помощью общедоступной программы JellyFish v.2.3.0 [125]. JellyFish представляет собой эффективный инструмент для подсчета подпоследовательностей (k-меров) заданной длины, в данном случае равной 23 нуклеотидам. Созданную базу 23-меров можно использовать для проверки прочтений на наличие контаминации. Результатом работы KrATER являются графики, представленные в двойном логарифмическом масштабе. По оси Y отложено число различающихся 23-меров, а по оси X кратность их покрытия. Обрезка адаптеров и фильтрация данных по качеству прочтений проводилась в два этапа: с первоначальной обрезкой на основе k-меров больших фрагментов адаптера с помощью Cookiecutter, а затем дополнительная обрезка малых фрагментов и фильтрация с помощью Trimmomatic v.0.36 с параметрами «SLIDINGWINDOW:8:20 MINLEN:50» [126,127]. Фильтрация каждого прочтения проходила скользящим окном, размер которого составлял 8 нуклеотидов. Если среднее качество прочтения в текущем окне падало ниже 20, происходило удаление оставшейся части прочтения. Если длина сохраненного участка составляла меньше 50 нуклеотидов - прочтение полностью удалялось. Целостность исследуемых геномных сборок была оценена с помощью программы QUAST v.5.0.2 [128]. Оценка качества геномных сборок по консервативным ортологичным группам была проведена с помощью BUSCO v.3 с использованием базы данных OrthoDB v.9 для млекопитающих (*Mammalia*) [129,130]. BUSCO предоставляет количественные показатели для оценки сборки генома, набора генов и полноты транскриптома, которые основаны на эволюционно-обоснованных ожиданиях содержания генов от почти универсальных ортологов, собранных в базе OrthoDB.

Выравнивание и поиск генетических вариантов

Сборка псевдохромосомного уровня на основе сцепленных прочтений 10X Genomics Польской популяции *Mustela nivalis* и референса *Mustela erminea* (https://www.ncbi.nlm.nih.gov/assembly/GCA_009829155.1), была получена с использованием инструмента RaGOO v.1.1, основанного на программе Minimap2 v.2.18 [131]. Выравнивание фильтрованных прочтений с соответствующими сборками генома *M. nivalis* проводилось с помощью инструмента Burrows-Wheeler Alignment (BWA) v.0.7.11 [132]. Дубликаты прочтений помечались с помощью программного пакета Samtools v.1.9 [133]. Поиск

генетических вариантов проводился с помощью Bcftools v.1.10 со следующими параметрами: «-d 250 -q 30 -Q 30 --adjust-MQ 50 -a AD,INFO/AD,ADF,INFO/ADF,ADR,INFO/ADR,DP,SP,SCR,INFO/SCR» для bcftools mpileup и «-m -O u -v -f GQ,GP» для bcftools call. Генетические варианты низкого качества ('QUAL < 20.0 || (FORMAT/SP > 60.0 || FORMAT/DP < 5.0 || FORMAT/GQ < 20.0)') были удалены с помощью bcftools filter [134]. Подсчет покрытия генома был выполнен с помощью программы Mosdepth v.0.3.0 [135]. Полученные результаты оценки покрытия визуализированы на тепловых картах с помощью набора программных инструментов MACE, основанных на библиотеке Matplotlib 2.

Визуализация гетерозиготности

Фильтрованные генетические варианты были разделены на однонуклеотидные полиморфизмы (SNP) и инсерции/делеции (indel). Все последующие анализы основывались только на данных SNP. Инделлы в последующем анализе не использовались из-за низкого качества при использовании коротких прочтений. Количество гетерозиготных SNP рассчитывалось в неперекрывающихся окнах размером 100 тыс п.н и 1 млн п.н. Полученные результаты были визуализированы на тепловых картах и боксплотах с помощью набора программных инструментов MACE, основанных на библиотеке Matplotlib 2.

Получение и визуализация сетей гаплотипов и филогенетического дерева

Для получения последовательности митохондриальной ДНК для каждого исследуемого образца, была использована программа MitoZ v.2.3 [136]. Полученные сборки мтДНК были использованы для выявления митохондриальных локусов CR (control region) и Cytb с помощью BLAST. Определение ядерного локуса ASIP было проведено посредством получения сборок из фильтрованных vcf файлов с помощью GATK v.4 для каждого образца с последующим поиском в полученных геномных сборках при помощи BLAST [137]. Определение гаплотипов CR, Cytb и ASIP из исследуемых образцов, проводилось с помощью программы CD-HIT v.4.8.1 [138]. Построение сетей гаплотипов было проведено в программе POPART v.1.7 с использованием модели «Median Joining networks» [139]. Филогенетическое дерево было построено с помощью программы MrBayes v.3.2.6, со следующими параметрами: модель HKY и GTR, 2000000 генераций MCMC, 4 цепи, 25% деревьев были отброшены как отжигаемая часть [140]. Полученное дерево было визуализировано при помощи FigTree v.1.4.2.

Реконструкция демографической истории

Реконструкция демографической истории была проведена с использованием программного пакета PSMC, восстанавливающего историю эффективного размера популяции на основе диплоидной последовательности, используя модели парного последовательного марковского слияния (PSMC) [103]. Создание консенсусной диплоидной последовательности было проведено с использованием программных инструментов Samtools v.0.1.19 [133]. Конвертация vcf файлов в fastq файлы проводилась с помощью vcfutils.pl с указанием следующих параметров минимального и максимального покрытия для vcf2fq: -d 12 -D 90 для образца 10X, -d 5 -D 35 для образца MNIV, -d 4 -D 32.5 для образца T100. Все, что находилось за пределами указанных диапазонов было отфильтровано. Для бутстреп репликации, инструментами PSMC, последовательности были разделены на более короткие сегменты. Бутстреп репликация рассчитывалась на 100 раундов с указанием параметра -b для каждого образца. Подготовка файлов к визуализации полученных результатов проводилась стандартными инструментами PSMC. Для параметра -g, указывающий на время генерации в годах, были переданы следующие значения: 1, 1.5, 2. Скорость замен (-u) использовалась стандартная - 2.5e-08. Для визуализация демографических историй был использован Google Sheets v.1.2.

Результаты и обсуждение

Горноста́й (*Mustela erminea*) - самый близкий к обыкновенной ласке (*Mustela nivalis*) вид с опубликованной геномной сборкой хромосомного уровня, а также по числу хромосом в кариотипе. У *M. nivalis* диплоидный набор включает $2n = 42$ хромосомы, а у *M. erminea* - $2n = 44$. По этим причинам черновая сборка обыкновенной ласки, полученная на основе сцепленных прочтений 10X Genomics Польской популяции была выровнена на референсную сборку хромосомного уровня *Mustela erminea*. В результате была получена псевдохромосомная сборка, при этом одна пара хромосом оказалась разорвана из-за отличий в числе хромосом этих видов. Получение сборки псевдохромосомного уровня упрощает визуализацию результатов анализа гетерозиготности. При решении других задач была использована исходная сборка на основе сцепленных прочтений 10X Genomics Польской популяции.

Оценка качества и фильтрация данных

Помня о принципе «Мусор на входе - мусор на выходе», для качественного проведения анализа и получения достоверных результатов, необходим обстоятельный подход к оценке качества данных, на которых строится все исследование. Учитывая все возможные ограничения в подходах, которые использовались для получения исходных данных, была проведена оценка качества и фильтрация. Для удаления из данных возможных технических последовательностей (адаптеры, индексы, баркоды) и данных низкого качества, была проведена фильтрация исходных данных (табл. 2). По результатам оценки качества фильтрованных данных, для референсного образца (10X) в процентном соотношении осталось меньше данных, чем для других двух образцов - MNIV и T100 (табл. 2). Важно отметить, что мы использовали достаточно жесткие параметры для фильтрации данных по качеству (см. Материалы и методы). Полученный результат для референсного образца не вызывает опасений об оставшемся количестве данных из-за исходно высокого покрытия, т.к. данных после фильтрации более чем достаточно для их анализа. Что касается использованной в работе *de novo* сборки, то геномный сборщик Supernova 2 был разработан в расчете на работу с сырыми данными секвенирования, и на этапе предобработки входных данных также проводит их фильтрацию. Таким образом, факт несколько меньшего качества входных данных не должен был существенно повлиять на процесс сборки, что подтверждается метриками качества. Нужно отметить, что собственно *de novo* сборка не

входит в данную работу и была проведена Jose Melo-Ferreira (Universidade do Porto, Португалия).

Важным этапом контроля качества является оценка различных метрик исходных данных. В случае геномных сборок наиболее критична метрика оценки целостности сборки. Одним из связанных с ней показателей является параметр N50, соответствующий размеру контига или скаффолда, который представляет собой наибольшую длину, начиная с которой в контигах такой же или большей длины содержится не менее половины всей сборки генома. В нашем случае даже у исходной черновой сборки N50 составляет ~25 млн п.н., что уже достаточно близко к хромосомному уровню, хотя и не достигает его. При таком значении N50, крупнейшие скаффолды сборки включают целые плечи хромосом, либо их крупные фрагменты. После скаффолдинга по геному горностая, целостность полученной псевдохромосомной сборки оказалась в ~5 раз выше (N50 ~130 млн п.н.), чем у исходной черновой сборки на основе сцепленных прочтений (табл. 3). Кроме того, суммарное число скаффолдов уменьшилось в пять раз, а количество С-скаффолдов в финальной сборке совпадает с таковым в использованной сборке генома горностая, что говорит об успешном скаффолдинге до хромосомного уровня.

Необходимо отметить, что сам по себе параметр N50 далеко не полностью отражает целостность и качество сборки. По этой причине мы также подсчитали общую длину пробелов в сборке и количество нуклеотидов, которые не удалось определить (Ns), а также определили минимальное количество скаффолдов, в которых содержится половина сборки (L50). Согласно L50, половина псевдохромосомной сборки содержится в восьми С-скаффолдах, что согласуется с кариотипическими данными, как у обыкновенной ласки, так и у горностая, так как в этих геномах присутствует большое число мелких хромосом [141].

Таблица 2. Характеристика использованных в работе геномных библиотек обыкновенной ласки.

№	ID образца*	Тип библиотеки	Количество пар прочтений до фильтрации, млн	Количество пар прочтений после фильтрации, млн
1	10X	ДНК	805.2	618.3
2	MNIV	ДНК	167.7	162.7
3	T100	ДНК	193.8	190.6

*Данные секвенирования геномов: 10X - сцепленные прочтения 10X Genomics (Польская популяция), MNIV и T100 - парные библиотеки Illumina 150bp (Новосибирская и Дальневосточная популяции).

Таблица 3. Характеристики исходной и псевдохромосомной сборок генома обыкновенной ласки.

№	Тип сборки*	Количество скаффолдов	Размер сборки, млрд п.н.	Ns, млн п.н.	N50, млн п.н.	L50
1	10X	33264	2.5	23.2	24.9	29
2	pChr	6735	2.5	25.8	130.1	8

*Сборки геномов: 10X - сборка генома на основе сцепленных прочтений 10X Genomics, pChr - сборка псевдохромосомного уровня.

На втором этапе контроля качества мы оценили целостность сборок по присутствию консервативных однокопийных ортологов млекопитающих из базы OrthoDB. Анализ статуса данных генов в исходной и псевдохромосомной сборках, проведенный при помощи BUSCO, дополняет статистические значения, рассмотренные выше (N50 и другие), и предоставляет количественные показатели для оценки качества сборок на основе эволюционно-обоснованных ожиданий содержания. При этом гены подразделяются на три категории - цельные (с двумя подкатегориями: однокопийные и дублицированные), фрагментированные и отсутствующие. Согласно результатам анализа, псевдохромосомная сборка незначительно превосходит исходную во всех категориях: в ней выше количество как цельных однокопийных генов, так и цельных генов в общем, а число отсутствующих, фрагментированных и дублицированных генов - ниже (табл. 4). При пересечении наборов генов по категориям между сборкам мы выявили интересный паттерн. Псевдохромосомная сборка была получена из исходной путем скаффолдинга, и, в идеальном случае, не должна была бы потерять при этом информацию (за исключением несхлопнутой гетерозиготности). Однако, мы видим, что присутствует небольшое количество генов со статусом «Отсутствующие» и «Фрагментированные» исключительно в псевдохромосомной сборке, в то время как в исходной сборке они имеют другой статус (рис. 11). Это может быть связано с артефактам как аннотации этих генов, так и скаффолдинга. При этом число таких генов невелико, а подавляющее большинство генов (~94% от всех проанализированных генов) были аннотированы целиком в обеих сборках, что говорит о достаточном качестве для анализа генов даже у исходной сборки на основе только сцепленных прочтений.

Таблица 4. Абсолютные значения оценки качества геномных сборок по консервативным ортологичным группам (BUSCOs).

Статус	Тип сборки*			
	pChr		10X	
	абс.	%	абс.	%
Цельные	3932	95.81%	3912	95.32%
- Цельные и однокопийные	3881	94.57%	3857	93.98%
- Цельные и дублицированные	51	1.24%	55	1.34%
Фрагментированные	85	2.07%	94	2.29%
Отсутствующие	87	2.12%	98	2.39%
Всего	4104		4104	

*Тип сборки: pChr - сборка псевдохромосомного уровня; 10X - сборка на основе сцепленных прочтений 10X Genomics.

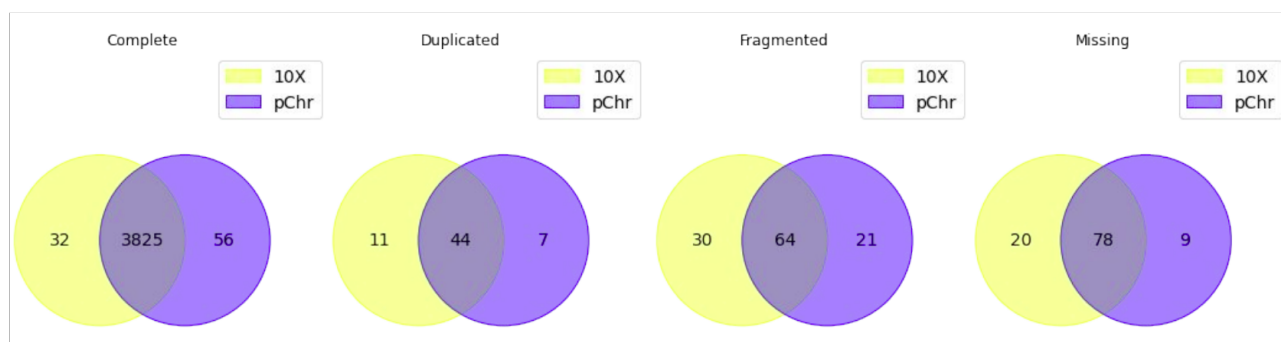


Рисунок 11. Перекрывание результатов оценки качества геномных сборок по различным категориям BUSCO.

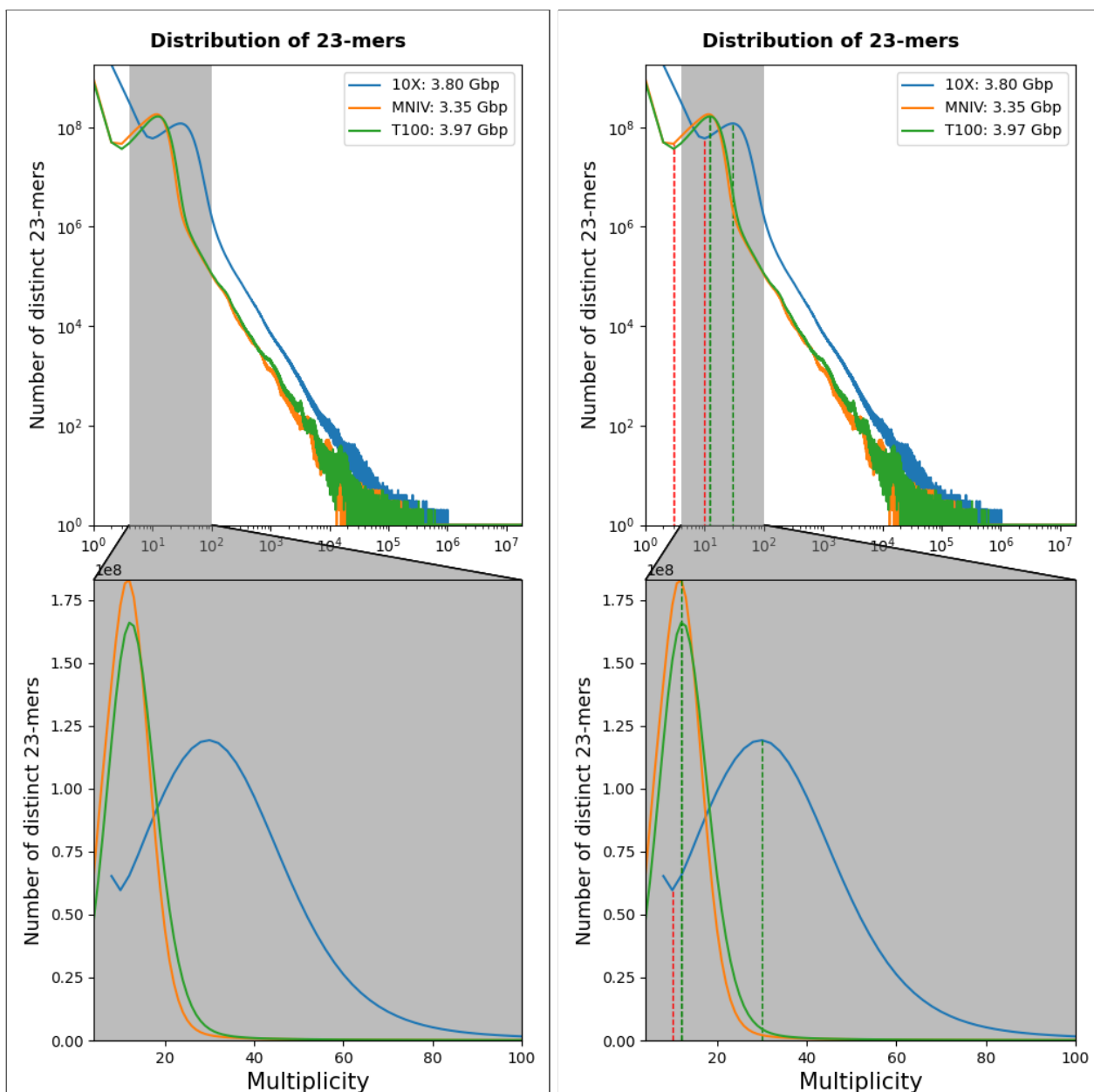


Рисунок 12. Контроль качества данных и оценка размера генома по распределению 23-меров для всех образцов: 10X - сцепленные прочтения 10X Genomics; MNIV и T100 - парные прочтения Illumina 150 п.н. Верхние графики представлены в логарифмическом масштабе по обеим осям, нижние - в линейном. По оси Y отложено число различающихся 23-меров, а по оси X - кратность их покрытия.

Еще одним важным этапом оценки качества данных, является анализ распределения k-меров - подпоследовательностей определенной длины k, в данном случае длина равна 23, - это стандартная длина, которая чуть выше того порога, когда в геноме появляется достаточно уникальных k-меров. Если образец не является сверхвысокогетерозиготным и в данных

отсутствуют протяженные дубликации и мультипликации, а также если нет контаминации чужеродной ДНК, то на графике будет присутствовать только один четкий пик.

По результатам анализа распределения 23-меров во всех трех образцах мы обнаружили по одному пику (рис. 12). Однако, в образце 10X ширина пика существенно больше, чем в двух других. Данный результат может быть объяснен отличием в покрытии образцов: для образца 10X медианное покрытие равно 36, для MNIV и T100 - 15 и 14 соответственно. Другой причиной может быть большая неравномерность покрытия в данном образце. Если сопоставить размер сборки (табл. 3) и размер генома, оцененный по распределению 23-меров (рис. 12), то видна существенная разница. В первом случае, размер сборки на основе сцепленных прочтений 10X Genomics составляет 2.5 млрд п.н., а по распределению 23-меров из тех же данных размер генома достигает 3.8 млрд п.н. Данные результаты могут быть объяснены присутствием в геноме больших гетерохроматиновых блоков, которые очень плохо собираются и часто отсутствуют в сборках геномов. Вариацию размера генома между образцами также можно объяснить разницей в размере и количестве гетерохроматиновых блоков в геномах особей из различных популяций. В остальном, никаких неожиданных особенностей в распределении 23-меров мы не выявили.

Визуализация покрытия и поиск псевдоаутосомного района (PAR)

После проведения оценки качества и фильтрации исходных данных, мы выровняли фильтрованные данные на псевдохромосомную сборку и визуализировали покрытие для всех трех образцов (рис. 13). Кратность покрытия аутосом во всех образцах составляет от 0.75x до 1.25x относительно медианного покрытия всего генома. Половые хромосомы образцов MNIV и 10X, имеют половинное покрытие (от 0.25x до 0.75x), что связано с полом этих образцов - оба образца являются самцами. Для образца T100 половая хромосома имеет по всей длине одно и то же покрытие, что и аутосомы, что говорит о половой принадлежности этого образца - образец является самкой. Первичное определения пола особей проводилось фенотипически при отлове. Первоначально, образцы были определены как две самки (MNIV и T100) и один самец (10X). Только в результате визуализации покрытия, образец MNIV был определен как самец, что впоследствии также было подтверждено нашими коллегами из ИМКБ СО РАН, посредством цитогенетического метода *in-situ* гибридизации (FISH) с использованием X хромосомы собаки (*Canis lupus*). Ошибка при первичном определении пола могла быть связана с редукцией первичных половых признаков (семенников) у особи MNIV.

Для указания ploидности при поиске генетических вариантов мы вычислили координаты псевдоаутосомных районов (PAR) по покрытию в сборке псевдохромосомного уровня (рис. 13). Псевдоаутосомный район представляет собой хорошо выраженный гомологичный участок половой хромосомы (X и Y в случае млекопитающих), частота рекомбинации в котором намного выше, чем в других геномных районах. PAR по покрытию X хромосомы можно выявить только у самцов, т.к. для них наблюдается половинное покрытие всей половой хромосомы, за исключением PAR, имеющего кратность покрытия намного выше. Псевдоаутосомный район занимает участок в X хромосоме от 0.07 млн п.н. до 6.55 млн п.н., и его координаты совпадают для обоих самцов, представленных образцами MNIV и 10X (рис. 13).

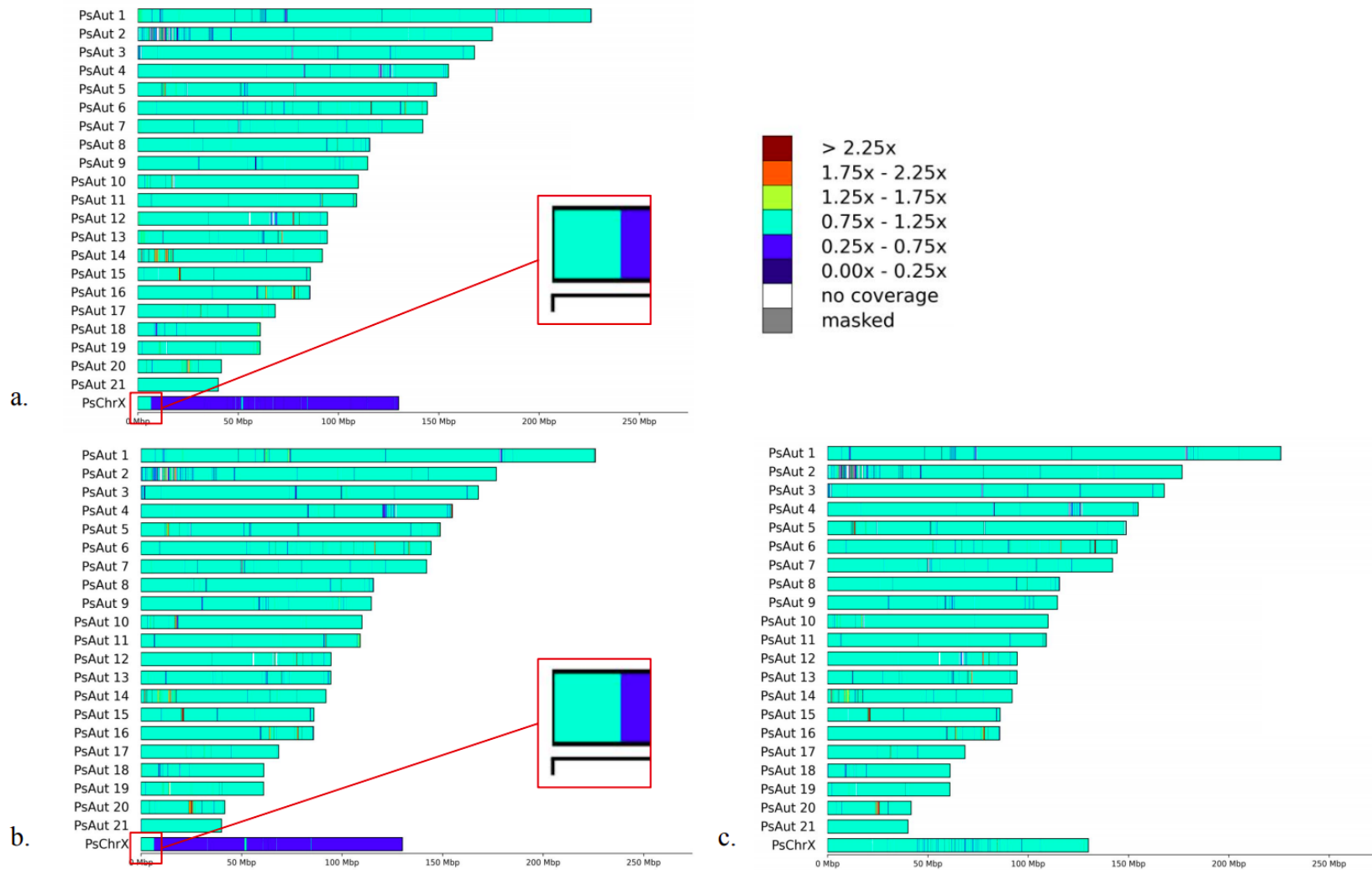


Рисунок 13. Результаты поиска псевдоаутосомного района (PAR) по покрытию в сборке псевдохромосомного уровня для всех трех образцов: MNIV (a), 10X (b), T100 (c). Красным квадратом обозначен псевдоаутосомный район, с координатами от 0.07 млн п.н. до 6.55 млн п.н.

Поиск генетических вариантов и визуализация гетерозиготности

Определение точных координат псевдоаутосомного района необходима для того, чтобы при поиске генетических вариантов задать ploidy для гаплоидной части X хромосомы у самцов. Первичную оценку гетерозиготности проще всего получить рассчитав среднюю гетерозиготность для всего генома, но использование такого подхода дает на выходе лишь одно число, которое не отражает распределение гетерозиготных локусов по геному. Более информативным способом является расчет средней и/или медианной гетерозиготности в разбивке на стыкующиеся или скользящие окна фиксированного размера. Выбор размера окна зависит от целостности сборки и плана проводимого анализа, но обычно выбранные размеры попадают в диапазон от 50 до 5.000 тыс п.н. При этом, значительная часть генома должна быть представлена в выбранных окнах, чтобы полученные результаты оценки гетерозиготности могли считаться достоверными. В случае геномных млекопитающих с типичным размером 2.5-4.0 млрд п.н. число окон размером 100 тыс п.н. обычно превышает >200.000 для сборок высокого качества, чего более чем достаточно для статистического анализа. Но для визуализации плотности гетерозиготности в масштабе всего генома, удобнее использовать более крупные окна длиной, например, в 1 млн п.н. С другой стороны для этого необходимо использовать сборки хромосомного уровня, чтобы избежать проблем с отображением большого количества скаффолдов на одном рисунке. По этой причине мы и провели описанный выше скаффолдинг исходной сборки ласки по референсному геному горностая.

В ранее опубликованных работах по изучению *Mustela nivalis* данные о степени генетического разнообразия на полногеномном уровне отсутствуют, но по результатам визуализации данных из нашего исследования, оказалось, что *Mustela nivalis* имеет высокую степень генетического разнообразия (рис. 14). При этом уровень гетерозиготности оцененный по SNP (рис. 14 a,b,c) и инделам (рис. 14 d,e,f) отличается кардинально. Данный факт можно объяснить двумя причинами. Точность выявления инделов в целом существенно ниже, чем SNP. Принципиально невозможно обнаружить инделы длиннее половины прочтения, то есть более 75 п.н. в нашем случае. Кроме того, частота появления инсерций и делеций сильно ниже, чем у SNP, что также приводит к меньшему их числу - как наблюдаемому, так и реальному. Несмотря на различия в средней гетерозиготности между образцами (рис. 15), для всех образцов присутствуют регионы с низким разнообразием (синий и темно-синий регионы на рис. 14). Такие протяженные участки с низкой плотностью

или даже отсутствием гетерозиготных SNP, отчетливо видны для образца T100 (рис. 14b), что может говорить о недавнем инбридинге.

При сравнении образцов между собой, наибольший уровень гетерозиготности продемонстрировал образец 10X, а наименьший - T100, соответственно (рис. 15). Образец MNIV при этом занимает промежуточное положение. Этот факт может быть связан с относительно невысоким покрытием образцов MNIV и T100 (14-15x), из-за которого некоторая часть гетерозиготных сайтов могла быть ошибочно принята за гомозиготные. Данное соотношение между образцами остается неизменным и при уменьшении размера окна с 1 млн п.н. (рис. 15b) до 100 тыс п.н. (рис. 15a).

Стоит отметить, что X хромосому можно определять не только по оценке покрытия у самцов, но и по плотности гетерозиготности у самок, по крайней мере у средне и высокогетерозиготных видов. Это может показаться парадоксальным, но регион X хромосомы гаплоидный у самцов, у самок имеет существенно более низкий уровень гетерозиготности по сравнению с аутосомами, чем ожидается даже с учетом меньшей численности X хромосом в популяции (0.75 относительно аутосом) [142]. В псевдоаутосомном районе, наоборот, генетическое разнообразие и у самцов и у самок всегда несколько выше, чем в аутосомах, т.к. в мейозе у хромосом X и Y присутствует одна облигатная хиазма, которая формируется всегда в псевдоаутосомном районе, как единственном гомологичном участке половых хромосом. Оба вышеописанных паттерна присутствуют в наших данных и четко видны на примере образца T100, который был получен от самки ласки (рис. 14b).

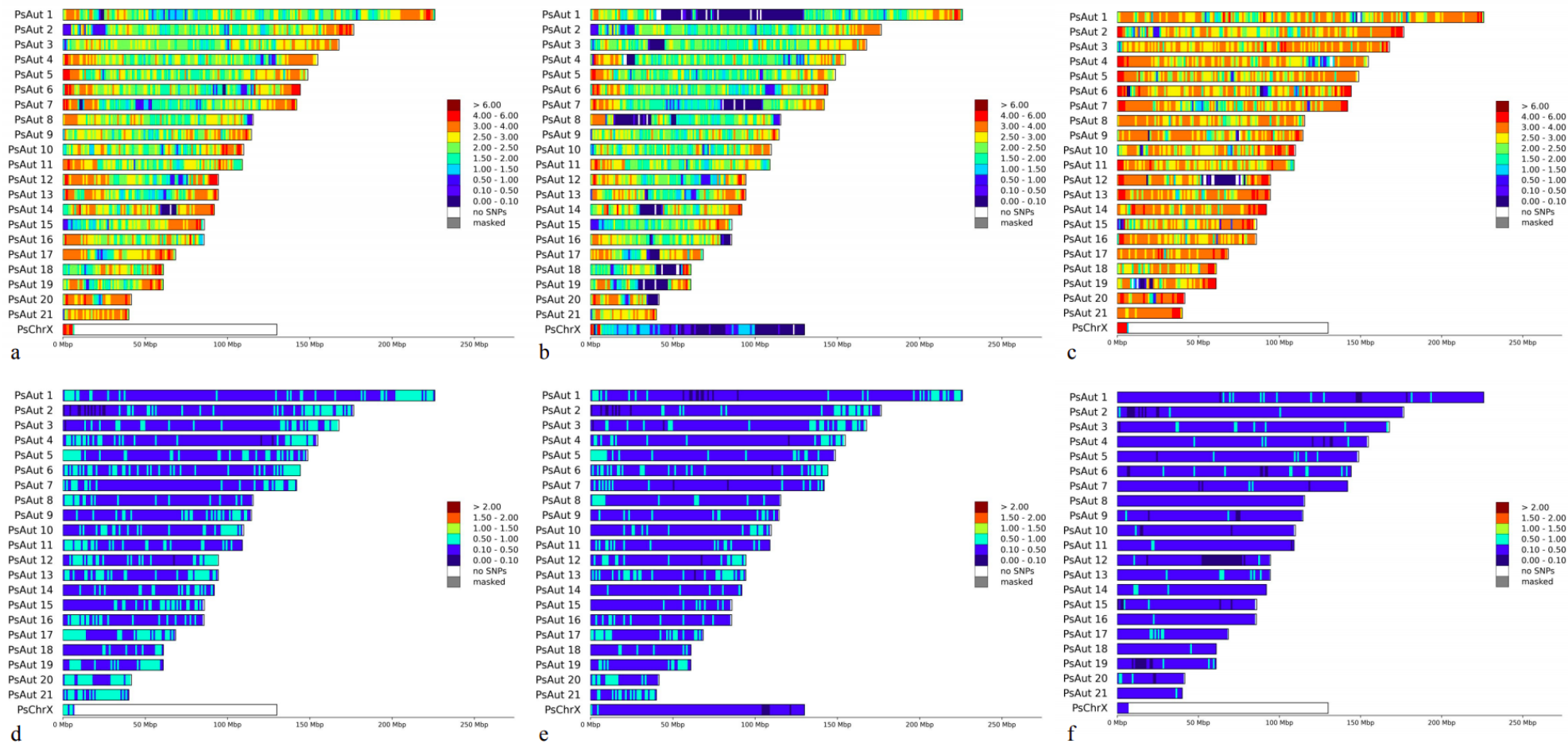


Рисунок 14. Распределение плотности гетерозиготных SNP (a, b, c) и инделов (d, e, f) в сборке псевдохромосомного уровня для всех 3 образцов: a, d - MNIV; b, e - T100; c, f - 10X. Количество гетерозиготных сайтов было подсчитано в стыкующихся окнах длиной в 1 млн п.н. и пересчитано в число сайтов на тыс п.н. При визуализации плотности SNP и инделов были использованы различные шкалы.

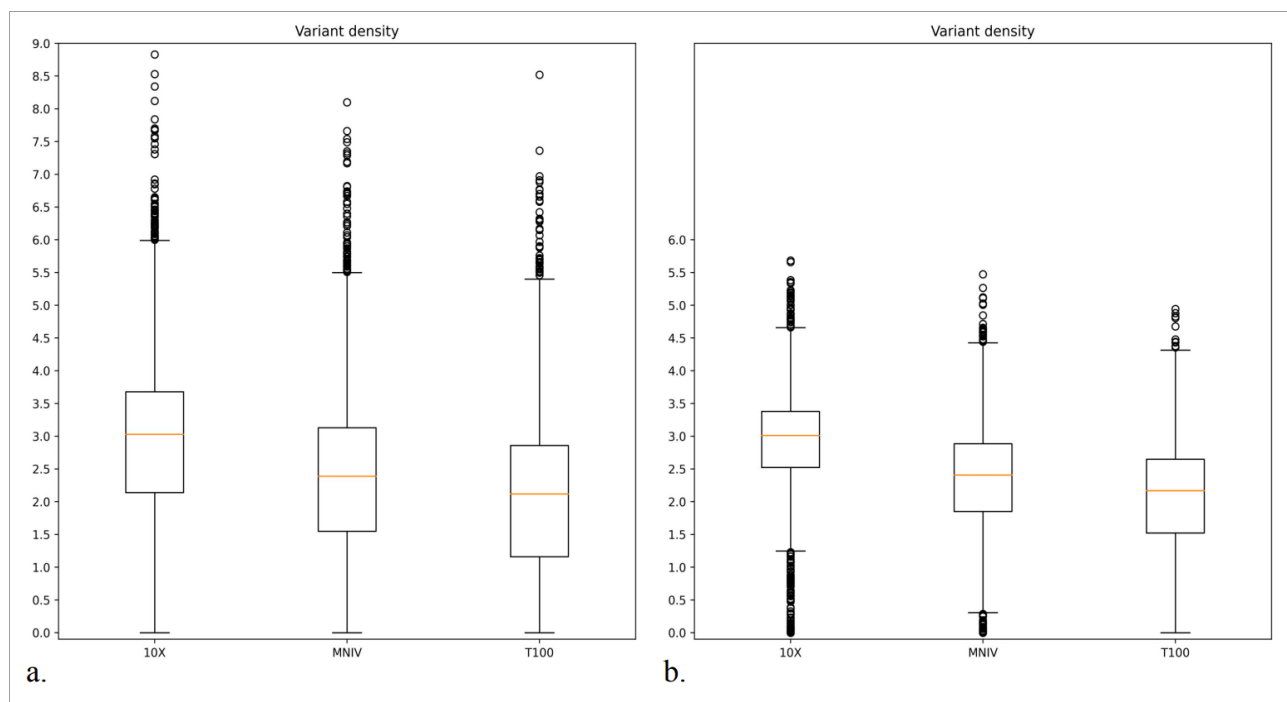


Рисунок 15. Сравнение распределения среднего значения гетерозиготности в окнах на 100Kbp (a) и 1Mbp (b) для сборки псевдохромосомного уровня.

Получение и визуализация сетей гаплотипов и построение филогенетического дерева

Полногеномных исследований *Mustela nivalis* ранее никто не проводил. Однако, в 2020 году была опубликована работа по изучению генетического разнообразия на основе митохондриальной ДНК и отдельных ядерных маркеров [11]. В результате работы были получены сети гаплотипов для относительно большого количества образцов *M. nivalis* из различных точек ареала. Мы сопоставили результаты исследования Sato et al. (2020) с образцами из нашей работы и определили, что конкатенированные данные CR (control region) и Cytb образцов MNIV и 10X совпадают с гаплотипами, обозначенными в работе Sato et al. (2020) как C35 и C36 соответственно, а для образца T100, конкатенированные данные не совпали ни с одним гаплотипом, поэтому он был обозначен нами как C120 (табл. 5). Гаплотипы ядерного маркера ASIP образцов из данного исследования соответствовали гаплотипу E6 из исследования Sato et al. (2020).

По полученным результатам, гаплотип C120, принадлежащий образцу T100, отличается от близлежащих гаплотипов C78 (Россия, п.Кедровая) и C1 (Южная Корея) всего одной заменой (рис. 16a). Поскольку образцы из Корейской и Дальневосточной популяций *M. nivalis* мало представлены в исследовании Sato et al. (2020), то возможно, что все три гаплотипа (C1, C78, C120) встречаются в обеих популяциях. Важно отметить, что гаплотип Новосибирского образца MNIV совпал с гаплотипом C35, который в работе Sato et al. (2020)

был выявлен у Польского образца. Такой результат можно объяснить смешанной структурой Польской популяции, которая предположительно образовалась в результате смешения различных предковых популяций [9]. Выбор ядерного локуса ASIP является не очень удачным, поскольку существенную часть его составляет экзон соответствующего гена, а на внутривидовом уровне из ядерного генома желательно выбирать вариабельные интроны. По результату построения сети гаплотипов ядерного локуса ASIP видно, что мы имеем достаточно однородную суперпопуляцию от Украины до Финляндии и до Дальнего востока (рис. 16b), причем все три образца из нашей работы имеют гаплотип E6.

По результату построения филогенетического дерева на конкатенированных данных митохондриальных локусов CR и Cytb (рис.17), мы определили, что наши образцы относятся к первой кладе, которая представлена в исследовании Sato et al. (2020) особями из Польши, от Испании до Южного Урала, Финляндии, Южной Кореи и Северного Урала. На построенном филогенетическом дереве были оставлены все ноды, даже с низкой постериорной вероятностью (<70%), которые не нужно учитывать, т.к. вероятность их существования очень низкая. Большое количество таких нод говорит о том, что в пределах первой и второй клад, выделение кластеров на филогенетическом дереве достаточно условное, в том числе и в исследовании Sato et al. (2020).

Таблица 5. Описание гаплотипов.

№	Образец*	Гаплотип			Источник гаплотипа из исследования Sato et al. (2020)	Идентификатор
		CR	Cytb	CR & Cytb		
1	MNIV	Po43	Po43	Po43	Польша, д.Беловежа	C35
2	T100	YP1	HS1119	**	Южная Корея (YP1), Россия, п.Кедровая (HS1119)	C120***
3	10X	Po44	Po44	Po44	Польша, д.Беловежа	C36

**Образцы из данной работы: 10X - сцепленные прочтения 10X Genomics (Польская популяция), MNIV и T100 - парные библиотеки Illumina 150 п.н. из (Новосибирская и Дальневосточная популяции).*

***В исследовании Sato et al. (2020) гаплотип отсутствует.*

****Впервые описан в данной работе.*

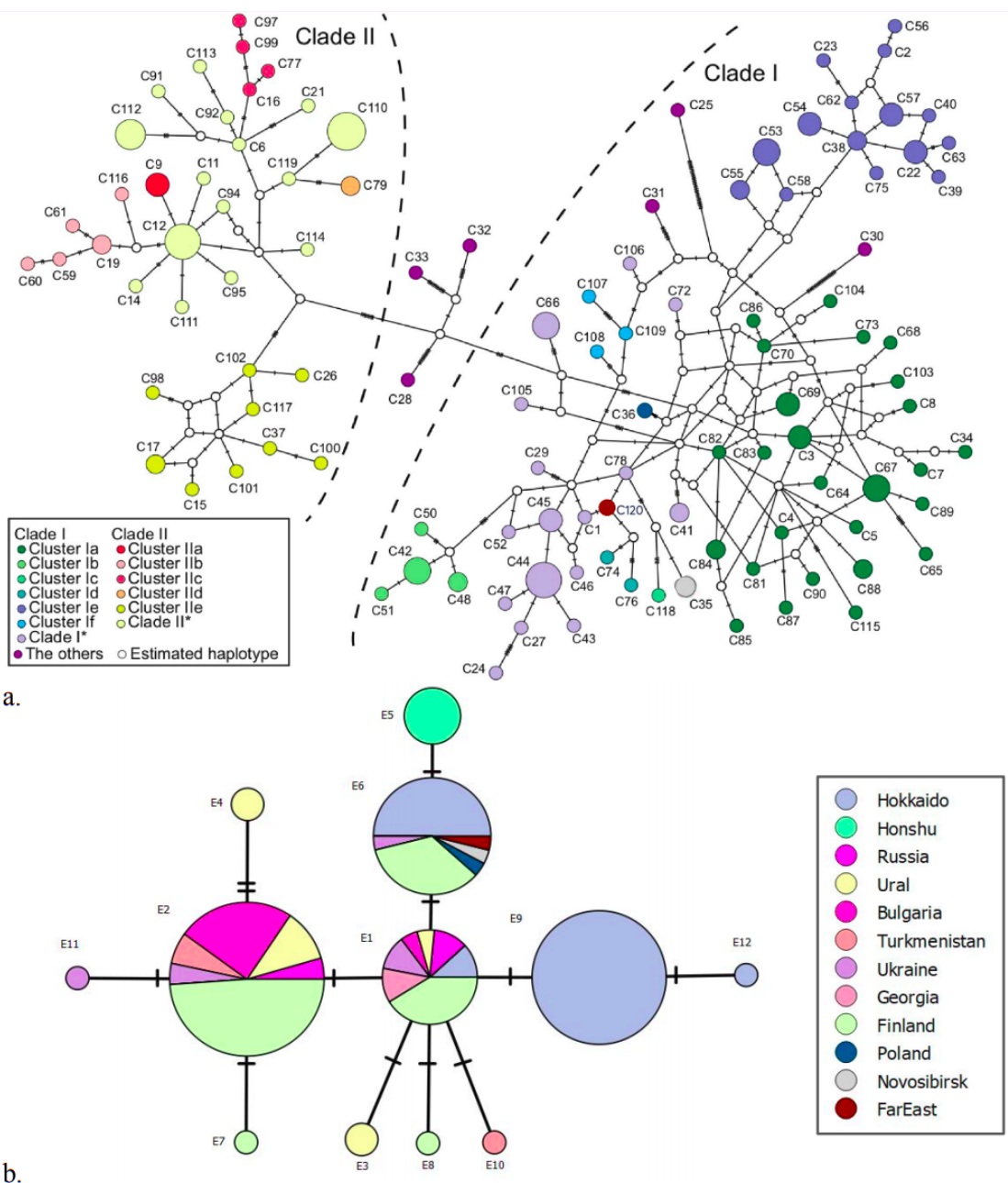


Рисунок 16. Сеть гаплотипов конкатенированных CR и Cytb мтДНК (а) и гаплотипов ядерного маркера ASIP (б). Рисунки из исследования Sato et al (2020) с добавлением образцов из данного исследования. Размеры кружков соответствуют частотам гаплотипов. Белые кружки обозначают предполагаемые гаплотипы, а количество штрихов на ребрах обозначает количество нуклеотидных замен между гаплотипами. Все образцы были разделены на две клады, оставшиеся образцы не входили ни в одну кладу. Звездочки означают, что гаплотипы не были сгруппированы в какие-либо кластеры, но включены в соответствующую кладу. Если нуклеотидные последовательности гаплотипов отличались только инделами или неразрешенными нуклеотидами, они рассматривались как один и тот же гаплотип.

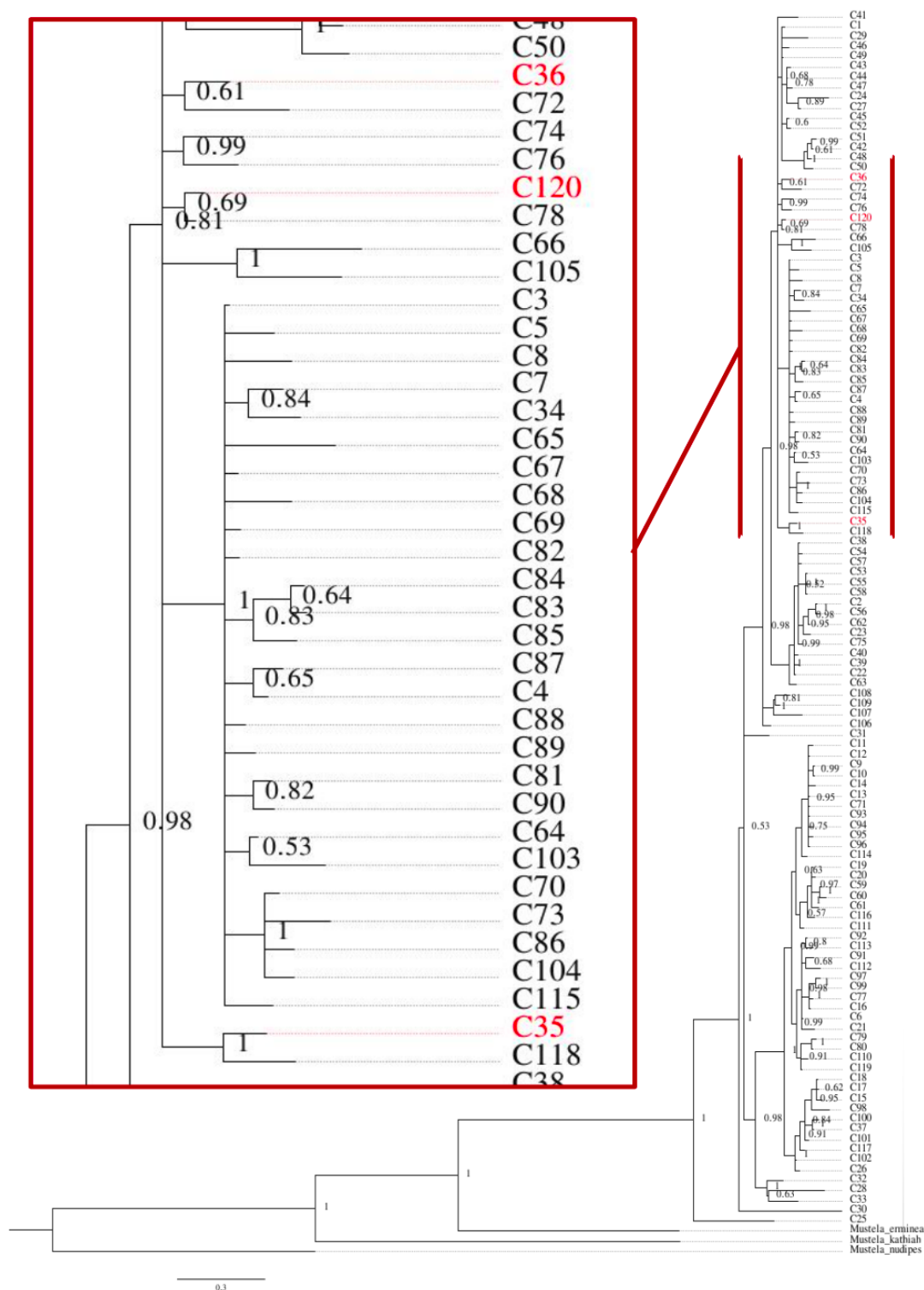


Рисунок 17. Байесовское филогенетическое дерево, построенное на конкатенированных данных митохондриальных локусов CR и Cytb. Добавленные образцы из данного исследования выделенные красным: C35 - MNIV (Новосибирская популяция), C36 - 10X (Польская популяция), C120 - T100 (Дальневосточная популяция). На дереве для нод обозначены постериорные вероятности. Ноды с низкой постериорной вероятностью были сохранены для иллюстрации условности выделения каких-либо кластеров в пределах клад.

Реконструкция демографической истории

Для реконструкции демографической истории с использованием метода PSMC, рекомендуется исключение из данных всех гомозиготных участков. По выявленному количеству и длине гомозиготных участков при визуализации гетерозиготности образцов (рис. 14), видно, что образцы Новосибирской и Дальневосточной популяций имеют отличия в их количестве. Тем не менее, полученный результат реконструкции демографической истории с сохранением всех имеющихся участков генома, показал практически идентичную демографическую динамику для образцов MNIV и T100 (рис. 18). Это позволяет предположить, что наличие гомозиготных участков в исследуемых образцах не оказало критического влияния на полученные результаты реконструкции демографической истории, а имеющиеся отличия находятся в пределах погрешности. Устойчивость полученных результатов к возмущениям во входных данных была проверена при помощи 100 бутстреп репликаторов для каждого из образцов.

Результаты реконструкции свидетельствуют об отделении предковой Польской популяции от остальных, за которым начался быстрый рост численности ~200.000 лет назад, что особенно заметно для предковой Польской популяции (рис. 18). Предковые популяции Новосибирска и Дальнего востока показали практически идентичную демографическую динамику, что позволяет предположить общее происхождение и/или постоянный и интенсивный поток генов. После отделения от предковой популяции, численность Сибирской популяции (MNIV и T100) оставалась стабильной в течение ~100.000 лет. Умеренный рост начался лишь ~80.000 лет назад, но тенденция изменилась ~30.000 лет назад. Таким образом, учитывая, что мы проанализировали лишь три образца, можно предположить, что резкая экспансия численности произошла лишь в Европе, а в Сибири численность обыкновенной ласки росла намного медленнее, причем наблюдается постепенное снижение численности, начавшееся ~30.000 лет назад. Одной из возможных причин тенденции снижения численности для Сибирских популяций, могут быть и различные межвидовые взаимодействия. Например, другой представитель семейства куньих - соболь (*Martes zibellina*), вытесняет из своего ареала более мелких представителей семейства, поскольку занимает схожую экологическую нишу, а также активно охотится на них. Лесная и каменная куница (*Martes martes* и *Martes foina*), распространенные в Европе, хоть и схожи с соболем, но занимают немного другую экологическую нишу из-за предпочтений к древесному образу жизни. Таким образом, различия в экологических взаимодействиях и конкурирующих видах могли повлиять на траекторию демографической истории.

При реконструкции демографической истории, рекомендуется исключать из данных половые хромосомы из-за возможного получения других частот генетических замен, что связано с отличием полового наследования от аутосомного. Численность X хромосом в популяции составляет 0.75, за исключением наличия псевдоаутосомного района. У PAR численность идентична с аутосомной, но в псевдоаутосомном районе происходит хиазма в процессе кроссинговера, что приводит к высокому генетическому разнообразию и скорости замен, по сравнению с половой хромосомой и даже аутосомами. Результат реконструкции демографической истории с исключением из данных половых хромосом, не выявил сильных отличий (рис. 19), которые не могли бы быть объяснены погрешностью реконструкции. Такой результат может быть связан с отсутствием в демографической истории исследуемых популяций *M. nivalis* довольно резких бутылочных горлышек численности с последующей резкой экспансией. Сравнение полученных результатов реконструкции демографической истории на основе полногеномных данных с половыми хромосомами (рис. 18) и без (рис. 19), показало что во втором случае траектории Новосибирской и Дальневосточной популяций сблизились по сравнению с первым случаем и практически совпадают, что дополнительно говорит в пользу гипотезы об однородной Сибирской суперпопуляции.

При анализе 100 бутстреп репликаций на полногеномных данных без половой хромосомы (рис. 19), мы выявили раздвоение пучка траекторий у Польской популяции в интервале от 60 до 300 тысяч лет назад, и аналогичное, но более слабое разделение у Дальневосточной популяции между 200.000 и 400.000 лет назад (рис. 19). Подобное разделение может означать присутствие двух локальных минимумов в реконструкции, что соответствует двум вариантам построения демографической истории Польской и Дальневосточной популяций в соответствующем временном интервале. Альтернативным объяснением может быть недостаточное количество бутстреп репликаций, либо нехватка данных. Предположение с двумя вариантами демографической истории соответствует наличию двух популяций или субпопуляций, которые впоследствии образовали одну. В пользу этого говорит, что раздвоение пучка траекторий совпадает с увеличением численности в случае и Польской и Дальневосточной популяций.

Однако, необходимо отметить, что высказанные выше гипотезы остаются в рамках спекуляции. Чтобы поставить точку в данном вопросе необходимо собрать и проанализировать большее количество образцов по всей Евразии, а также использовать альтернативные методы реконструкции демографической истории.

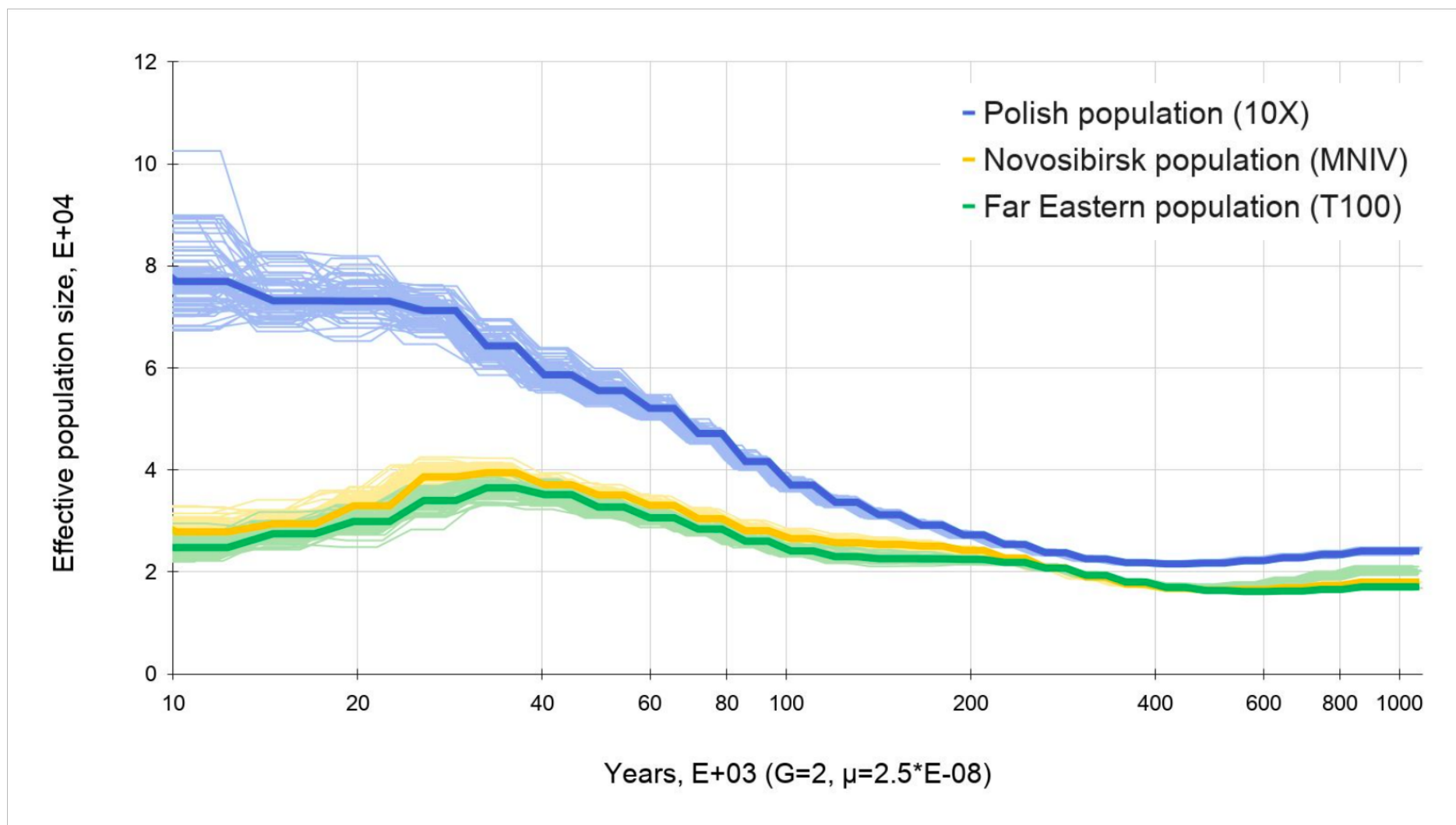


Рисунок 18. Реконструкция демографической истории трех популяций *Mustela nivalis* на основе полногеномных данных.

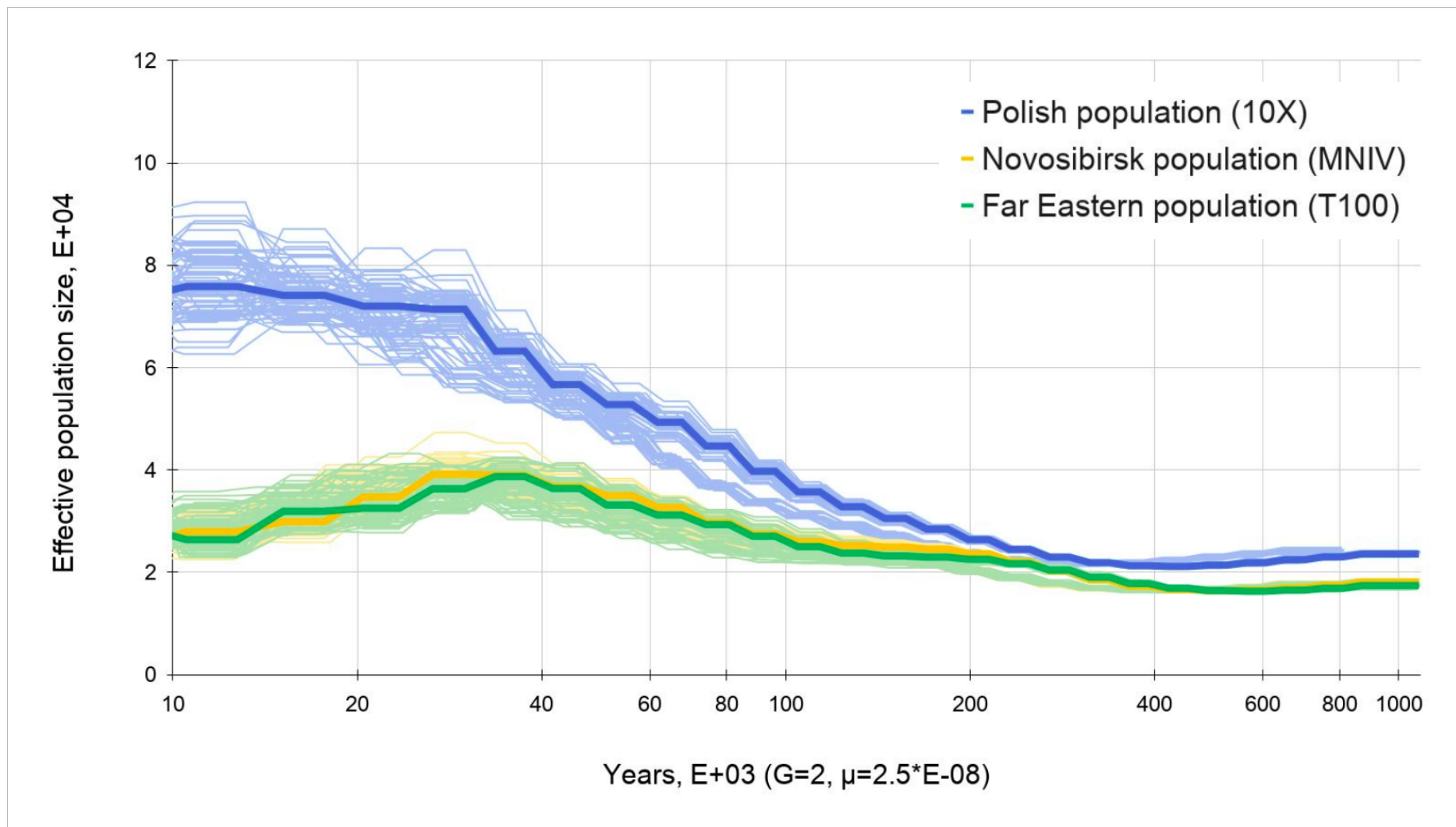


Рисунок 19. Реконструкция демографической истории трех популяций *Mustela nivalis* на основе полногеномных данных без половых хромосом.

Выводы

1. Данные о демографической истории трех условных популяций подтверждают ранние исследования об относительно быстром распространении *M. nivalis* по Палеарктике, которое могло произойти в результате постледниковой реколонизации из различных рефугиумов, которые характеризуются смешением популяций.
2. Выявленная идентичная демографическая динамика популяций Новосибирска и Дальнего востока, вероятно, свидетельствует о реколонизации из одного и того же рефугиума и/или интенсивном потоке генов, что привело к формированию однородной Сибирской суперпопуляции, несмотря на отдаленное расположение этих популяций, расстояние между которыми составляет тысячи километров.
3. Ранее неизвестный высокий уровень генетического разнообразия *M. nivalis*, в Дальневосточной популяции сочетается с присутствием протяженных гомозиготных участков, что может говорить о недавнем инбридинге в этой популяции.
4. Распределение гетерозиготных вариантов в X хромосоме демонстрирует характерные для нее паттерны, которые можно использовать для локализации псевдоаутосомного района у высокогетерозиготных видов используя данные самок, а не самцов.
5. Для подтверждения высказанных предположений необходимо провести масштабный анализ демографической истории *M. nivalis* на полногеномном уровне, представив популяции Евразии в полной мере.

Благодарности

Выражаю особую благодарность моему научному консультанту н.с. ИМКБ СО РАН Кливеру Сергею Федоровичу, за профессиональную помощь и неоценимую поддержку на протяжении всех этапов исследования, благодаря которой появилась уверенность в своих силах.

Благодарю всех преподавателей Биологического факультета СПбГУ и, в частности, с.н.с., к.б.н. Райко Михаила Петровича за высокий профессионализм, компетентность и возможность выполнения данной работы.

Благодарю Полину Перельман, Наталью Сердюкову, Виолетту Р. Беклемишеву, Татьяну М. Бульонкову, Inês Miranda и Liliana Farelo за выделение ДНК и создание клеточных культур; Karol Zub, Виктора В. Панова, Анну С. Мухачеву и Алексея В. Абрамова - за сбор образцов и предоставление данных; Александра С. Графодатского, Klaus-Peter Koepli и Jose Melo-Ferreira - за финансирование и организацию работы; Алексея Сергушичева и Владимира Ульянцева - за организацию работы на вычислительном кластере КТ ИТМО.

Работа была поддержана финансированием со стороны JetBrains Research и грантом РФФИ 20-04-00808.

Список литературы

1. Ceballos G., Ehrlich P.R., Dirzo R. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines // Proc. Natl. Acad. Sci. National Academy of Sciences, 2017. Vol. 114, № 30. P. E6089–E6096.
2. Fuentes-Pardo A.P., Ruzzante D.E. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations // Mol. Ecol. 2017. Vol. 26, № 20. P. 5369–5406.
3. Spielman D., Brook B.W., Frankham R. Most species are not driven to extinction before genetic factors impact them // Proc. Natl. Acad. Sci. National Academy of Sciences, 2004. Vol. 101, № 42. P. 15261–15264.
4. Frankham R. et al. Introduction to Conservation Genetics. Cambridge University Press, 2002. 642 p.
5. Morikawa M.K., Palumbi S.R. Using naturally occurring climate resilient corals to construct bleaching-resistant nurseries // Proc. Natl. Acad. Sci. National Academy of Sciences, 2019. Vol. 116, № 21. P. 10586–10591.
6. Hoban S. et al. Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved // Biol. Conserv. 2020. Vol. 248. P. 108654.
7. Miller W. et al. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change // Proc. Natl. Acad. Sci. National Academy of Sciences, 2012. Vol. 109, № 36. P. E2382–E2390.
8. Abascal F. et al. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx // Genome Biol. 2016. Vol. 17, № 1. P. 251.
9. Abramov A.V., Baryshnikov G.F. Geographic variation and intraspecific taxonomy of weasel *Mustela nivalis* (Carnivora, *Mustelidae*). Vol. 8. P. 38.
10. McDevitt A.D. et al. Climate and refugial origin influence the mitochondrial lineage distribution of weasels (*Mustela nivalis*) in a phylogeographic suture zone // Biol. J. Linn. Soc. 2012. Vol. 106, № 1. P. 57–69.
11. Sato T. et al. Phylogeography and population history of the least weasel (*Mustela nivalis*) in the Palearctic based on multilocus analysis // J. Zool. Syst. Evol. Res. 2020. Vol. 58, № 1. P. 408–426.
12. Douglas J. Biologists urge US endowment for conservation // Nature. 1978. Vol. 275. P. 82–83.
13. Mares M.A. Review of Conservation Biology: An Evolutionary-Ecological Perspective //

- Wilson Bull. Wilson Ornithological Society, 1981. Vol. 93, № 3. P. 420–422.
14. Cooke S.J. et al. What is conservation physiology? Perspectives on an increasingly integrated and essential science† // *Conserv. Physiol.* 2013. Vol. 1, № cot001.
 15. Margules C.R., Pressey R.L. Systematic conservation planning: 6783 // *Nature*. Nature Publishing Group, 2000. Vol. 405, № 6783. P. 243–253.
 16. Braverman I. Conservation without nature: the trouble with in situ versus ex situ conservation // *Geoforum*. 2014. Vol. 51. P. 47–57.
 17. Raup D.M. A Kill Curve For Phanerozoic Marine Species // *Paleobiology*. Paleontological Society, 1991. Vol. 17, № 1. P. 37–48.
 18. Ceballos G. et al. Accelerated modern human-induced species losses: Entering the sixth mass extinction // *Sci. Adv.* American Association for the Advancement of Science, 2015. Vol. 1, № 5. P. e1400253.
 19. The IUCN Red List of Threatened Species [Electronic resource] // IUCN Red List of Threatened Species. URL: <https://www.iucnredlist.org/en> (accessed: 07.03.2021).
 20. Baumel A. R. Frankham, J. D. Ballou & D. A. Briscoe, *Introduction to Conservation Genetics*, Cambridge University Press, Cambridge (2002) // *Ecol. Mediterr. Persée - Portail des revues scientifiques en SHS*, 2003. Vol. 29, № 1. P. 116–117.
 21. Gaston K.J. Valuing Common Species // *Science*. American Association for the Advancement of Science, 2010. Vol. 327, № 5962. P. 154–155.
 22. Vié J.-C., Hilton-Taylor C., Stuart S.N. *Wildlife in a Changing World: An Analysis of the 2008 IUCN Red List of Threatened Species*. IUCN, 2009. 184 p.
 23. Díaz S. et al. Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. 2019.
 24. Blanchet S. et al. A river runs through it: The causes, consequences, and management of intraspecific diversity in river networks // *Evol. Appl.* 2020. Vol. 13, № 6. P. 1195–1213.
 25. Allendorf F.W. Genetics and the conservation of natural populations: allozymes to genomes // *Mol. Ecol.* 2017. Vol. 26, № 2. P. 420–430.
 26. Hohenlohe P.A., Funk W.C., Rajora O.P. Population genomics for wildlife conservation and management // *Mol. Ecol.* 2021. Vol. 30, № 1. P. 62–82.
 27. Thaden A. von et al. Applying genomic data in wildlife monitoring: Development guidelines for genotyping degraded samples with reduced single nucleotide polymorphism panels // *Mol. Ecol. Resour.* 2020. Vol. 20, № 3. P. 662–680.
 28. Höglund J. et al. Phylogeography of willow grouse (*Lagopus lagopus*) in the Arctic:

- taxonomic discordance as inferred from molecular data // Biol. J. Linn. Soc. 2013. Vol. 110, № 1. P. 77–90.
29. Milano I. et al. Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (*Merluccius merluccius*) // Mol. Ecol. 2014. Vol. 23, № 1. P. 118–135.
 30. McMahon B.J., Teeling E.C., Höglund J. How and why should we implement genomics into conservation? // Evol. Appl. 2014. Vol. 7, № 9. P. 999–1007.
 31. Zhan X. et al. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle: 5 // Nat. Genet. Nature Publishing Group, 2013. Vol. 45, № 5. P. 563–566.
 32. Fabbri E. et al. From the Apennines to the Alps: colonization genetics of the naturally expanding Italian wolf (*Canis lupus*) population // Mol. Ecol. 2007. Vol. 16, № 8. P. 1661–1671.
 33. Holliday J.A., Hallerman E.M., Haak D.C. Genotyping and Sequencing Technologies in Population Genetics and Genomics // Population Genomics: Concepts, Approaches and Applications / ed. Rajora O.P. Cham: Springer International Publishing, 2019. P. 83–125.
 34. Meek M.H., Larson W.A. The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era // Mol. Ecol. Resour. 2019. Vol. 19, № 4. P. 795–803.
 35. Andrews K.R. et al. Harnessing the power of RADseq for ecological and evolutionary genomics: 2 // Nat. Rev. Genet. Nature Publishing Group, 2016. Vol. 17, № 2. P. 81–92.
 36. Hohenlohe P.A. et al. Population Genomics Provides Key Insights in Ecology and Evolution // Population Genomics: Concepts, Approaches and Applications / ed. Rajora O.P. Cham: Springer International Publishing, 2019. P. 483–510.
 37. Garner B.A. et al. Genomics in Conservation: Case Studies and Bridging the Gap between Data and Application // Trends Ecol. Evol. Elsevier, 2016. Vol. 31, № 2. P. 81–83.
 38. Walters A.D., Schwartz M.K. Population Genomics for the Management of Wild Vertebrate Populations // Population Genomics: Wildlife / ed. Hohenlohe P.A., Rajora O.P. Cham: Springer International Publishing, 2021. P. 419–436.
 39. Funk W.C. et al. Harnessing genomics for delineating conservation units // Trends Ecol. Evol. 2012. Vol. 27, № 9. P. 489–496.
 40. Saremi N.F. et al. Puma genomes from North and South America provide insights into the genomic consequences of inbreeding: 1 // Nat. Commun. Nature Publishing Group, 2019. Vol. 10, № 1. P. 4769.
 41. Förster D.W. et al. Targeted resequencing of coding DNA sequences for SNP discovery in nonmodel species // Mol. Ecol. Resour. John Wiley & Sons, Ltd, 2018. Vol. 18, № 6. P.

- 1356–1373.
42. Meek M.H. et al. Sequencing improves our ability to study threatened migratory species: Genetic population assignment in California's Central Valley Chinook salmon // *Ecol. Evol.* 2016. Vol. 6, № 21. P. 7706–7716.
 43. Davis C.S., Strobeck C. Isolation, variability, and cross-species amplification of polymorphic microsatellite loci in the family *Mustelidae* // *Mol. Ecol.* 1998. Vol. 7, № 12. P. 1776–1778.
 44. Li B. et al. Genetic individualization of sable (*Martes zibellina* L. 1758) using microsatellites // *Anim. Cells Syst.* Taylor & Francis, 2018. Vol. 22, № 4. P. 253–258.
 45. Brandies P. et al. The Value of Reference Genomes in the Conservation of Threatened Species: 11 // *Genes*. Multidisciplinary Digital Publishing Institute, 2019. Vol. 10, № 11. P. 846.
 46. Janecka J.E. et al. Noninvasive Genetics and Genomics Shed Light on the Status, Phylogeography, and Evolution of the Elusive Snow Leopard // *Conservation Genetics in Mammals: Integrative Research Using Novel Approaches* / ed. Ortega J., Maldonado J.E. Cham: Springer International Publishing, 2020. P. 83–120.
 47. Clendenin H.R. et al. Combining Harvest and Genetics to Estimate Reproduction in Wolves // *J. Wildl. Manag.* 2020. Vol. 84, № 3. P. 492–504.
 48. Charlesworth B. Effective population size and patterns of molecular evolution and variation: 3 // *Nat. Rev. Genet.* Nature Publishing Group, 2009. Vol. 10, № 3. P. 195–205.
 49. Tautz D. Hypervariability of simple sequences as a general source for polymorphic DNA markers // *Nucleic Acids Res.* 1989. Vol. 17, № 16. P. 6463–6471.
 50. Baron E.E. et al. Parentage testing and effect of misidentification on the estimation of breeding value in Gir cattle // *Genet. Mol. Biol.* Sociedade Brasileira de Genética, 2002. Vol. 25, № 4. P. 389–394.
 51. Bishop M.D. et al. A genetic linkage map for cattle. // *Genetics*. Genetics, 1994. Vol. 136, № 2. P. 619–639.
 52. Mburu D., Hanotte O. A practical approach to microsatellite genotyping with special reference to livestock population genetics. P. 82.
 53. Goldstein D., Pollock D. Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. // *J. Hered.* 1997.
 54. Beaumont M.A. Detecting Population Expansion and Decline Using Microsatellites // *Genetics*. 1999. Vol. 153, № 4. P. 2013–2029.
 55. Charlesworth D., Willis J.H. The genetics of inbreeding depression: 11 // *Nat. Rev. Genet.* Nature Publishing Group, 2009. Vol. 10, № 11. P. 783–796.
 56. Allendorf F.W., Berry O., Ryman N. So long to genetic diversity, and thanks for all the fish //

- Mol. Ecol. 2014. Vol. 23, № 1. P. 23–25.
57. Hill W.G. Estimation of effective population size from data on linkage disequilibrium1 // Genet. Res. Cambridge University Press, 1981. Vol. 38, № 3. P. 209–216.
 58. Waples R.S., Do C. Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution // Evol. Appl. 2010. Vol. 3, № 3. P. 244–262.
 59. Luikart G. et al. Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches // Conserv. Genet. 2010. Vol. 11, № 2. P. 355–373.
 60. Li M.H., Merilä J. Extensive linkage disequilibrium in a wild bird population: 6 // Heredity. Nature Publishing Group, 2010. Vol. 104, № 6. P. 600–610.
 61. Edwards A. et al. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups // Genomics. 1992. Vol. 12, № 2. P. 241–253.
 62. Baird N.A. et al. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers // PLOS ONE. Public Library of Science, 2008. Vol. 3, № 10. P. e3376.
 63. Davey J.W., Blaxter M.L. RADSeq: next-generation population genetics // Brief. Funct. Genomics. 2010. Vol. 9, № 5–6. P. 416–423.
 64. Nunziata S.O., Weisrock D.W. Estimation of contemporary effective population size and population declines using RAD sequence data: 3 // Heredity. Nature Publishing Group, 2018. Vol. 120, № 3. P. 196–207.
 65. Lehnert S.J. et al. Genomic signatures and correlates of widespread population declines in salmon: 1 // Nat. Commun. Nature Publishing Group, 2019. Vol. 10, № 1. P. 2996.
 66. Grossen C. et al. Population genomics analyses of European ibex species show lower diversity and higher inbreeding in reintroduced populations // Evol. Appl. 2018. Vol. 11, № 2. P. 123–139.
 67. Schyler O. Nunziata et al. Genomic data detect corresponding signatures of population size change on an ecological time scale in two salamander species // Mol. Ecol. 2017. Vol. 26, № 4. P. 1060–1074.
 68. Jensen E.L. et al. Population genomics through time provides insights into the consequences of decline and rapid demographic recovery through head-starting in a Galapagos giant tortoise // Evol. Appl. 2018. Vol. 11, № 10. P. 1811–1821.
 69. Allendorf F.W., Hohenlohe P.A., Luikart G. Genomics and the future of conservation genetics: 10 // Nat. Rev. Genet. Nature Publishing Group, 2010. Vol. 11, № 10. P. 697–709.
 70. Li J. et al. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? // Mol. Ecol. 2012. Vol. 21, № 1. P. 28–44.

71. Stapley J. et al. Adaptation genomics: the next generation // Trends Ecol. Evol. 2010. Vol. 25, № 12. P. 705–712.
72. Davey J.W. et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing // Nat. Rev. Genet. 2011. Vol. 12, № 7. P. 499–510.
73. Bonneaud C. et al. Rapid evolution of disease resistance is accompanied by functional changes in gene expression in a wild bird // Proc. Natl. Acad. Sci. U. S. A. 2011. Vol. 108, № 19. P. 7866–7871.
74. Storz J.F. et al. Evolutionary and functional insights into the mechanism underlying high-altitude adaptation of deer mouse hemoglobin // Proc. Natl. Acad. Sci. U. S. A. 2009. Vol. 106, № 34. P. 14450–14455.
75. Hohenlohe P.A. et al. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags // PLoS Genet. 2010. Vol. 6, № 2. P. e1000862.
76. Parchman T.L. et al. The genomic consequences of adaptive divergence and reproductive isolation between species of manakins // Mol. Ecol. 2013. Vol. 22, № 12. P. 3304–3317.
77. Catchen J. et al. The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing // Mol. Ecol. 2013. Vol. 22, № 11. P. 2864–2883.
78. De Wit P., Palumbi S.R. Transcriptome-wide polymorphisms of red abalone (*Haliotis rufescens*) reveal patterns of gene flow and local adaptation // Mol. Ecol. 2013. Vol. 22, № 11. P. 2884–2897.
79. Reitzel A.M. et al. Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics // Mol. Ecol. 2013. Vol. 22, № 11. P. 2953–2970.
80. White T.A. et al. Adaptive evolution during an ongoing range expansion: the invasive bank vole (*Myodes glareolus*) in Ireland // Mol. Ecol. 2013. Vol. 22, № 11. P. 2971–2985.
81. Keller I. et al. Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes // Mol. Ecol. 2013. Vol. 22, № 11. P. 2848–2863.
82. Hohenlohe P.A. et al. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing // Mol. Ecol. 2013. Vol. 22, № 11. P. 3002–3013.
83. Alex Buerkle C., Gompert Z. Population genomics based on low coverage sequencing: how low should we go? // Mol. Ecol. 2013. Vol. 22, № 11. P. 3028–3035.
84. Seeb J.E. et al. Single-nucleotide polymorphism (SNP) discovery and applications of SNP

- genotyping in nonmodel organisms // *Mol. Ecol. Resour.* 2011. Vol. 11 Suppl 1. P. 1–8.
85. Hohenlohe P.A. et al. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout // *Mol. Ecol. Resour.* 2011. Vol. 11 Suppl 1. P. 117–122.
 86. Ogden R. et al. Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing // *Mol. Ecol.* 2013. Vol. 22, № 11. P. 3112–3123.
 87. Goodwin S., McPherson J.D., McCombie W.R. Coming of age: ten years of next-generation sequencing technologies: 6 // *Nat. Rev. Genet.* Nature Publishing Group, 2016. Vol. 17, № 6. P. 333–351.
 88. da Fonseca R.R. et al. Next-generation biology: Sequencing and data analysis approaches for non-model organisms // *Mar. Genomics.* 2016. Vol. 30. P. 3–13.
 89. Schlötterer C. et al. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding: 11 // *Nat. Rev. Genet.* Nature Publishing Group, 2014. Vol. 15, № 11. P. 749–763.
 90. Nagasaki M. et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals // *Nat. Commun.* 2015. Vol. 6. P. 8018.
 91. Snyder M.W. et al. Haplotype-resolved genome sequencing: experimental methods and applications // *Nat. Rev. Genet.* 2015. Vol. 16, № 6. P. 344–358.
 92. Raineri E. et al. SNP calling by sequencing pooled samples // *BMC Bioinformatics.* 2012. Vol. 13. P. 239.
 93. Gutenkunst R.N. et al. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data // *PLoS Genet.* 2009. Vol. 5, № 10. P. e1000695.
 94. Han E., Sinsheimer J.S., Novembre J. Fast and accurate site frequency spectrum estimation from low coverage sequence data // *Bioinforma. Oxf. Engl.* 2015. Vol. 31, № 5. P. 720–727.
 95. Ronen R. et al. Learning natural selection from the site frequency spectrum // *Genetics.* 2013. Vol. 195, № 1. P. 181–193.
 96. Lamichhaney S. et al. Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean // *Proc. Natl. Acad. Sci. U. S. A.* 2017. Vol. 114, № 17. P. E3452–E3461.
 97. Trame C.B. et al. New mini- zincin structures provide a minimal scaffold for members of this metallopeptidase superfamily // *BMC Bioinformatics.* 2014. Vol. 15. P. 1.
 98. Zhao S. et al. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation // *Nat. Genet.* 2013. Vol. 45, № 1. P. 67–71.
 99. Foote A.D. et al. Genome-culture coevolution promotes rapid divergence of killer whale

- ecotypes // Nat. Commun. 2016. Vol. 7. P. 11693.
100. Nadachowska-Brzyska K. et al. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula flycatchers* // Mol. Ecol. 2016. Vol. 25, № 5. P. 1058–1072.
 101. Jónás Á. et al. Estimating the Effective Population Size from Temporal Allele Frequency Changes in Experimental Evolution // Genetics. 2016. Vol. 204, № 2. P. 723–735.
 102. Vieira F.G., Albrechtsen A., Nielsen R. Estimating IBD tracts from low coverage NGS data // Bioinforma. Oxf. Engl. 2016. Vol. 32, № 14. P. 2096–2102.
 103. Li H., Durbin R. Inference of human population history from individual whole-genome sequences // Nature. 2011. Vol. 475, № 7357. P. 493–496.
 104. Schiffels S., Durbin R. Inferring human population size and separation history from multiple genome sequences // Nat. Genet. 2014. Vol. 46, № 8. P. 919–925.
 105. Spence J.P. et al. Inference of population history using coalescent HMMs: review and outlook // Curr. Opin. Genet. Dev. 2018. Vol. 53. P. 70–76.
 106. Orlando L. et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse // Nature. 2013. Vol. 499, № 7456. P. 74–78.
 107. Palkopoulou E. et al. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth // Curr. Biol. CB. 2015. Vol. 25, № 10. P. 1395–1400.
 108. Skoglund P. et al. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds // Curr. Biol. CB. 2015. Vol. 25, № 11. P. 1515–1519.
 109. Mather N., Traves S.M., Ho S.Y.W. A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data // Ecol. Evol. 2020. Vol. 10, № 1. P. 579–589.
 110. Ho S.Y.W., Shapiro B. Skyline-plot methods for estimating demographic history from nucleotide sequences // Mol. Ecol. Resour. 2011. Vol. 11, № 3. P. 423–434.
 111. Prado-Martinez J. et al. Great ape genetic diversity and population history // Nature. 2013. Vol. 499, № 7459. P. 471–475.
 112. Mazet O. et al. On the importance of being structured: instantaneous coalescence rates and human evolution--lessons for ancestral population size inference? // Heredity. 2016. Vol. 116, № 4. P. 362–371.
 113. Freedman A.H. et al. Genome sequencing highlights the dynamic early history of dogs // PLoS Genet. 2014. Vol. 10, № 1. P. e1004016.
 114. Terhorst J., Kamm J.A., Song Y.S. Robust and scalable inference of population history from hundreds of unphased whole genomes // Nat. Genet. 2017. Vol. 49, № 2. P. 303–309.

115. Weissman D.B., Hallatschek O. Minimal-assumption inference from population-genomic data // *eLife*. 2017. Vol. 6.
116. Ogilvie H.A., Bouckaert R.R., Drummond A.J. StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates // *Mol. Biol. Evol.* 2017. Vol. 34, № 8. P. 2101–2114.
117. Fu Q. et al. Genome sequence of a 45,000-year-old modern human from western Siberia // *Nature*. 2014. Vol. 514, № 7523. P. 445–449.
118. Cahill J.A. et al. Inferring species divergence times using pairwise sequential Markovian coalescent modelling and low-coverage genomic data // *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 2016. Vol. 371, № 1699.
119. Fan S. et al. African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations // *Genome Biol.* 2019. Vol. 20, № 1. P. 82.
120. Besenbacher S. et al. Direct estimation of mutations in great apes reconciles phylogenetic dating // *Nat. Ecol. Evol.* 2019. Vol. 3, № 2. P. 286–292.
121. Beichman A.C., Huerta-Sanchez E., Lohmueller K.E. Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms // *Annu. Rev. Ecol. Evol. Syst.* 2018. Vol. 49, № 1. P. 433–456.
122. Shajii A., Numanagić I., Berger B. Latent Variable Model for Aligning Barcoded Short-Reads Improves Downstream Analyses // *Res. Comput. Mol. Biol. Annu. Int. Conf. RECOMB Proc. RECOMB Conf. 2005-.* 2018. Vol. 10812. P. 280–282.
123. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom, 2010.
124. Kliver Sergei et al. KrATER: K-mer Analysis Tool Easy to Run. Институт проблем передачи информации им. А.А. Харкевича РАН, 2017. P. 128.
125. Marçais G., Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers // *Bioinformatics*. 2011. Vol. 27, № 6. P. 764–770.
126. Bolger A.M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data // *Bioinformatics*. 2014. Vol. 30, № 15. P. 2114–2120.
127. Starostina E. et al. Cookiecutter: a tool for kmer-based read filtering and extraction // *bioRxiv*. Cold Spring Harbor Laboratory, 2015. P. 024679.
128. Gurevich A. et al. QUAST: quality assessment tool for genome assemblies // *Bioinformatics*. 2013. Vol. 29, № 8. P. 1072–1075.
129. Kriventseva E.V. et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs // *Nucleic*

- Acids Res. 2019. Vol. 47, № D1. P. D807–D811.
130. Simão F.A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs // *Bioinformatics*. 2015. Vol. 31, № 19. P. 3210–3212.
 131. Alonge M. et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes // *Genome Biol*. 2019. Vol. 20, № 1. P. 224.
 132. Li H., Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform // *Bioinformatics*. 2009. Vol. 25, № 14. P. 1754–1760.
 133. Li H. et al. The Sequence Alignment/Map format and SAMtools // *Bioinformatics*. 2009. Vol. 25, № 16. P. 2078–2079.
 134. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data // *Bioinformatics*. 2011. Vol. 27, № 21. P. 2987–2993.
 135. Pedersen B.S., Quinlan A.R. Mosdepth: quick coverage calculation for genomes and exomes // *Bioinformatics*. 2018. Vol. 34, № 5. P. 867–868.
 136. Meng G. et al. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization // *Nucleic Acids Res*. 2019. Vol. 47, № 11. P. e63–e63.
 137. McKenna A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data // *Genome Res*. 2010. Vol. 20, № 9. P. 1297–1303.
 138. Fu L. et al. CD-HIT: accelerated for clustering the next-generation sequencing data // *Bioinformatics*. 2012. Vol. 28, № 23. P. 3150–3152.
 139. Leigh J.W., Bryant D. popart: full-feature software for haplotype network construction // *Methods Ecol. Evol*. 2015. Vol. 6, № 9. P. 1110–1116.
 140. Ronquist F. et al. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space // *Syst. Biol*. 2012. Vol. 61, № 3. P. 539–542.
 141. Graphodatsky A., Perelman P., O’Brien S.J. *Atlas of Mammalian Chromosomes*. John Wiley & Sons, Incorporated, 2020.
 142. Osada N. et al. Finding the Factors of Reduced Genetic Diversity on X Chromosomes of *Macaca fascicularis*: Male-Driven Evolution, Demography, and Natural Selection // *Genetics*. 2013. Vol. 195, № 3. P. 1027–1035.